# involve

## a journal of mathematics

**msp**

# 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.

# Finding cycles in the $k$-th power digraphs over the integers modulo a prime

Greg Dresden and Wenda Tu

(Communicated by Kenneth S. Berenhaut)

For $p$ prime and $k \geq 2$, let us define $G_p^{(k)}$ to be the digraph whose set of vertices is $\{0, 1, 2, \ldots, p-1\}$ such that there is a directed edge from a vertex $a$ to a vertex $b$ if $a^k \equiv b \bmod p$. We find a new way to decide if there is a cycle of a given length in a given graph $G_p^{(k)}$.

## Introduction

Let $k \geq 2$ be an integer and let $p$ be prime. Let us define $G_p^{(k)}$ to be the digraph whose set of vertices is $\{0, 1, 2, \ldots, p-1\}$ such that there is a directed edge from a vertex $a$ to a vertex $b$ if $a^k \equiv b \bmod p$.

This paper extends the results given in [Somer and Křížek 2004] (which provides a way to determine whether there is a cycle of length $t$ in a given graph $G_p^{(2)}$) and [Wilson 1998] (which considers $G_p^{(k)}$; see also [Somer and Křížek 2009]). In this paper, we provide our own way to determine the existence of a cycle of given length in $G_p^{(k)}$. First, we examine the existence of length-$t$ cycles where $t$ is prime. Later on, we explore the case of cycles of length $u$ where $u$ is composite, and we conclude with a study of digraphs that admit some cycle lengths but do not allow others.

Now, we will introduce one of the key theorems of this paper, mentioned in [Niven, Zuckerman and Montgomery 1991]. Here, $\phi$ stands for the Euler totient function.

**Theorem 1.** *Suppose $m = 1, 2, 4, p^\alpha$ or $2p^\alpha$, where $p$ is an odd prime and $\alpha$ is a positive integer. If $\gcd(a, m) = 1$ then the congruence $x^n \equiv a \bmod m$ has $\gcd(n, \phi(m))$ solutions or no solution, according to whether*

$$a^{\phi(m)/\gcd(n, \phi(m))} \equiv 1 \bmod m$$

*or not.*

---

## On the existence of length-$t$ cycles given $t$ prime, and length-$u$ cycles given $u \geq 2$

Based on the theorem in our Introduction, we have the following corollaries, which are crucial in determining the existence of a length-$t$ cycle for $t$ prime.

**Corollary 2.** *Let $p$ be a prime. The congruence $x^n \equiv 1$ mod $p$ has $\gcd(n, p-1)$ solutions.*

**Corollary 3.** *Let $p$ be a prime and let $k \geq 2$. The subgraph $G_p^{(k)} \setminus \{0\}$ has $\gcd(k-1, p-1)$ cycles of length $1$.*

Since we are curious about the existence of length-$t$ cycles in $G_p^{(k)}$ given $t$ prime, we want to know if the following equations have any solutions:

$$x^{k^t} \equiv x \mod p, \quad x^k \not\equiv x \mod p.$$

By our two corollaries, the above equations are equivalent to

$$\gcd(k^t - 1, p - 1) > \gcd(k - 1, p - 1).$$

Similarly, since we are also curious about the existence of length-$u$ cycles in $G_p^{(k)}$ given $u$ composite, we want to know if the following equations have any solutions (here, $u_i$ runs over the proper divisors of $u$):

$$x^{k^u} \equiv x \mod p, \quad x^{k^{u_1}} \not\equiv x \mod p, \quad x^{k^{u_2}} \not\equiv x \mod p, \quad \cdots.$$

Once again, our corollaries tell us that the above equations are equivalent to

$$\gcd(k^u - 1, p - 1) > \gcd(k^{u_i} - 1, p - 1)$$

for $u_i$ running over all proper divisors of $u$. So, we have the following results:

**Theorem 4.** *Given $u \geq 2$, $k \geq 2$, and $p$ prime, there exists a length-$u$ cycle in $G_p^{(k)}$ if and only if $\gcd(k^u - 1, p - 1) > \gcd(k^{u'} - 1, p - 1)$ for all proper divisors $u'$ of $u$.*

**Remark.** Theorem 4 follows from Theorem 5.6 of [Somer and Křížek 2009], which also gives formulas for how many cycles exist of a given length; if $t$ is prime, for example, then the number of length-$t$ cycles is

$$\frac{\gcd(k^t - 1, p - 1) - \gcd(k - 1, p - 1)}{t}.$$

The following theorem, which is a result of [Lucheta, Miller and Reiter 1996, pp. 230–231], is also a special case of a more general result in [Wilson 1998, pp. 232–233]. Another version with $k = 2$ appeared in [Somer and Křížek 2004, Theorem 3.3].

**Theorem 5** (Lucheta, Miller, Reiter). *Let $p$ be a prime. There exists a cycle of length $u$ in $G_p^{(k)}$ if and only if $u = \operatorname{ord}_d k$ for some divisor $d$ of $p - 1$ with $\gcd(d, k) = 1$, where $\operatorname{ord}_d k$ denotes the multiplicative order of $k$ modulo $d$.*

Here are four corollaries following from Theorems 4 and 5 that give us precise information on what cycle lengths are possible (or impossible) in $G_p^{(k)}$ for various primes $p$ and powers $k$:

**Corollary 6.** *Fix a prime $t$. Given any integer $k \geq 2$, there are infinitely many primes $p$ such that $G_p^{(k)}$ has a length-$t$ cycle. Moreover, $G_p^{(k)}$ contains a 1-cycle for all primes $p$.*

**Corollary 7.** *Fix an integer $u \geq 2$. Given any integer $k \geq 2$, there are infinitely many primes $p$ such that $G_p^{(k)}$ does not have a length-$u$ cycle.*

**Corollary 8.** *Fix an integer $u \geq 2$. Let $p = 2^{2^n} + 1$ be a Fermat prime, where $n \geq 0$. The possible cycle lengths in $G_p^{(k)}$ for $p$ a Fermat prime are very limited*:

(1) *There are never any odd-length cycles (aside from the length-1 cycles).*

(2) *If $k$ is even, there are no cycles at all (aside from the length-1 cycles) in $G_p^{(k)}$.*

(3) *If $k$ is odd and $u$ is even, $G_p^{(k)}$ contains a length-$u$ cycle if and only if $u \mid \mathrm{ord}_{p-1} k$. Moreover, $\mathrm{ord}_{p-1} k \mid 2^{2^n - 2}$ if $n \geq 2$ and $\mathrm{ord}_{p-1} k \mid 2^{2^n - 1}$ if $n = 0$ or 1.*

**Corollary 9.** *Fix an integer $u \geq 2$, and let $p$ be prime. Then, there are infinitely many integers $k$ such that $G_p^{(k)}$ contains no length-$u$ cycle.*

*Proof of Corollary 6.* Since $\gcd\left(1, \sum_{i=0}^{t-1} k^i\right) = 1$, by Dirichlet's theorem on the infinitude of primes in arithmetic progressions we know that there are infinitely many primes $p$ such that $p \equiv 1 \bmod \sum_{i=0}^{t-1} k^i$. Now given such a prime $p$, we have

$$\gcd\left((k-1)\sum_{i=0}^{t-1} k^i, \; p-1\right) \geq \sum_{i=0}^{t-1} k^i \quad \text{or} \quad \gcd(k^t - 1, \; p-1) \geq \sum_{i=0}^{t-1} k^i.$$

On the other hand, $\gcd(k-1, p-1) \leq k-1$. Since it is not hard to see that $k - 1 < \sum_{i=0}^{t-1} k^i$, we have $\gcd(k^t - 1, p-1) > \gcd(k-1, p-1)$. Thus, by Theorem 4, we can conclude that there are infinitely many primes $p$ such that $G_p^{(k)}$ has a length-$t$ cycle, as desired. Finally, the last assertion of our statement holds, since both 0 and 1 are clearly vertices in 1-cycles. □

*Proof of Corollary 7.* Let $q_1, q_2, \ldots$ be the odd primes in order of size, and let $q_r$ be the largest prime less than or equal to $k^u$; since both $u$ and $k$ are at least 2, then $q_r$ is at least 3. By the Chinese remainder theorem and Dirichlet's theorem, there exist infinitely many primes $p$ such that

$$p \equiv 3 \bmod 4, \quad p \equiv 2 \bmod q_1 q_2 \cdots q_r.$$

The first equivalence implies $2 \mid p - 1$ but $4 \nmid p - 1$, while the second implies $p - 1$ is relatively prime to $q_1 q_2 \cdots q_r$. Now suppose $G_p^{(k)}$ actually does have a length-$u$ cycle for $u \geq 2$. It follows from Theorem 5 that $u = \mathrm{ord}_d k$ for some divisor $d > 1$

of $p - 1$ (note that if $d = 1$ then this would imply $u = 1$, a contradiction), with $d$ relatively prime to $k$. Let us consider the options, keeping in mind what we just wrote about $p - 1$. If $k$ is odd, then either $d = 2$ or $d \geq q_{r+1}$. But if $d = 2$ then $u = \text{ord}_2 k = 1$, which is a contradiction. Hence, our only option is $d \geq q_{r+1}$. If $k$ is even, then $d$ must be odd and so again our only option is $d \geq q_{r+1}$. But with $d \geq q_{r+1}$, since $1 < k^u < q_{r+1}$ we have $u$ is not, in fact, the order of $k$ mod $d$, which contradicts our statement earlier that $u = \text{ord}_d k$. Hence, $G_p^{(k)}$ does not have a length-$u$ cycle for $u \geq 2$. $\square$

Before we move on to the next proof, we need to establish this useful result.

**Lemma 10.** *For $k$ odd and $a \geq 2$, the order $\text{ord}_{2^{a+1}} k$ is either equal to $\text{ord}_{2^a} k$ or to $2 \, \text{ord}_{2^a} k$.*

*Proof of Lemma 10.* If we let $w = \text{ord}_{2^a} k$, then we know that $2^a \mid k^w - 1$. Consider $k^{2w} - 1 = (k^w - 1)(k^w + 1)$. We know $2^a$ divides $k^w - 1$ and since $k$ is odd, 2 divides $k^w + 1$, so we know $2^{a+1}$ divides $k^{2w} - 1$. Hence, $\text{ord}_{2^{a+1}} k$ divides $2w$, but $\text{ord}_{2^{a+1}} k$ is at least $w$, and so we conclude that $\text{ord}_{2^{a+1}} k$ is either $w$ or $2w$, as desired. $\square$

We are now ready for the following:

*Proof of Corollary 8.* Let $p$ be the Fermat prime $2^{2^n} + 1$ where $n \geq 0$, so $p - 1 = 2^{2^n}$. Now, suppose $G_p^{(k)}$ contains a cycle of length $u \geq 2$. Then by Theorem 5, $u = \text{ord}_d k$ for some divisor $d$ of $p - 1 = 2^{2^n}$. By Euler's generalization of Fermat's little theorem, this implies $u \mid \phi(d)$, but $d$ is a power of 2 and so (thanks to the well-known formulas for Euler's phi function) this implies $u$ is as well. By Theorem 5, we also have $d$ and $k$ are relatively prime; since $u \mid d$, we know $u$ and $k$ are relatively prime as well. With this in mind, let us consider the possibilities for $u$ and $k$. We cannot have $u \geq 2$ be an odd integer, as this contradicts $u \geq 2$ being a power of 2; hence, $G_p^{(k)}$ never contains a cycle of length $u \geq 2$ for $u$ odd. We also cannot have $u$ and $k$ both be even integers, as this contradicts $u$ and $k$ being relatively prime; hence, $G_p^{(k)}$ contains no cycles of length $u$ for $u$ and $k$ both even.

The only option left is to have $u \geq 2$ even and $k \geq 2$ odd. Theorem 5 tells us that we have a length-$u$ cycle if and only if $u = \text{ord}_d k$ for some divisor $d$ of $p - 1$; let us establish that this is equivalent to $u \mid \text{ord}_{p-1} k$. For the first Fermat prime $p = 2^{2^0} + 1 = 3$, it is easy to verify that there are no even-length cycles in $G_3^{(k)}$ because this graph contains only the vertices $\{0, 1, 2\}$; likewise, $\text{ord}_{p-1} k = 1$ and this admits no even divisors. For the next Fermat prime $p = 2^{2^1} + 1 = 5$, similar calculations reveal that we can have even length-$u$ cycles only for $k \equiv 3 \bmod 4$ and for $u = 2$, in which case $u$ is indeed an even divisor of $2 = \text{ord}_{p-1} k$ (and vice versa). For both of those two cases (namely, for $p = 2^{2^n} + 1$ with $n = 0$ or 1), it is easy to check that $\text{ord}_{p-1} k \mid 2^{2^n - 1}$, as desired.

It remains to consider the other Fermat primes $p = 2^{2^n} + 1$ for $n \geq 2$. If $u = \text{ord}_d k$ for some divisor $d$ of $p - 1 = 2^{2^n}$, then (recalling that $u$ and $d$ and $p - 1$ are all

powers of 2) it is certainly true that $u \mid \operatorname{ord}_{p-1} k$, as $\operatorname{ord}_d k$ cannot be greater than $\operatorname{ord}_{p-1} k$ and both are powers of 2. For the other direction, suppose $u \mid \operatorname{ord}_{p-1} k$, and let us show that $u = \operatorname{ord}_d k$ for some divisor $d$ of $p - 1$. Starting with 1 as the order of $k$ mod 2, we imagine finding the orders of $k$ mod $2^2$, mod $2^3$, mod $2^4$, and so on, up to mod $2^{2^n}$. Lemma 10 tells us that at each step, the order of $k$ either stays the same or doubles. At the last step in this sequence (modulo $2^{2^n}$) the order of $k$ is a multiple of $u$. Hence, at some step along the way (say, when our modulus is $2^b$ for $b \leq 2^n$) we know that the order of $k$ mod $2^b$ is equal to $u$. Hence, we let $d = 2^b$ and we have $u = \operatorname{ord}_d k$ for $d$ a divisor of $p - 1$, as desired.

Finally, we recall from [Gallian 2010, p. 160] that the multiplicative group of units modulo $2^{2^n}$, commonly written $(\mathbb{Z}_{2^{2^n}})^*$, is isomorphic to $\mathbb{Z}_{2^{2^n-2}} \oplus \mathbb{Z}_2$ for $n \geq 2$. Hence, the order of any odd number $k$ modulo $p - 1$ will be a divisor of $2^{2^n-2}$, as desired. $\square$

*Proof of Corollary 9.* Note that if $p$ is a Fermat prime, then by Corollary 8 we can simply choose $k$ to be any even number. Of course, for $p = 2$ the conclusion is trivial. For the more general case, we choose $k \geq 2$ to be an integer equivalent to 1 mod $p - 1$. There are clearly infinitely many such $k$. Note that $\gcd(k, p-1) = 1$ and also $k \equiv 1$ mod $d$ for any divisor $d$ of $p - 1$. Thus, $\operatorname{ord}_d k = 1$ for any divisor $d$ of $p - 1$ and so by Theorem 5 we know $G_p^{(k)}$ has no $u$-cycles for any $u \geq 2$. $\square$

## On the existence of cycles of different lengths in the same digraph

We now consider cycles of composite length, and we show that the existence of certain cycles implies the existence of other, longer cycles.

**Theorem 11.** *Let $u = \operatorname{lcm}(u_1, u_2)$, where $u_1$ and $u_2$ are positive integers. If $G_p^{(k)}$ contains cycles of length $u_1$ and length $u_2$ respectively, then $G_p^{(k)}$ also contains a cycle of length $u$.*

*Proof.* Suppose in $G_p^{(k)}$ there exist cycles of lengths $u_1$ and $u_2$. By Theorem 5, we know that there exist $d_1$ and $d_2$ such that $d_1 \mid (p-1)$, $d_2 \mid (p-1)$ and $u_1 = \operatorname{ord}_{d_1} k$, $u_2 = \operatorname{ord}_{d_2} k$. Also, let $d = \operatorname{lcm}(d_1, d_2)$ and $u = \operatorname{lcm}(u_1, u_2)$. Since $d_1 \mid (k^{u_1} - 1)$ and $u_1 \mid u$, we have $d_1 \mid (k^u - 1)$. By the same reasoning, $d_2 \mid (k^u - 1)$. Therefore, $d \mid (k^u - 1)$; that is, $k^u \equiv 1$ mod $d$. So, $\gcd(d, k) = 1$. Assume there exists $u' \leq u$ such that $k^{u'} \equiv 1$ mod $d$. So, $k^{u'} \equiv 1$ mod $d_1$ and $k^{u'} \equiv 1$ mod $d_2$. Since $u_1$ is the order of $k$ mod $d_1$, we have $u_1 \mid u'$. Likewise, $u_2 \mid u'$. Therefore, $\operatorname{lcm}(u_1, u_2) \mid u'$; that is, $u \mid u'$. So $u \leq u'$. By assumption we know $u' \leq u$; thus, $u = u'$. So, the order of $k$ mod $d$ is $u$. Since $d = \operatorname{lcm}(d_1, d_2)$, we have $d \mid (p - 1)$. Thus again by Theorem 5, we know there is a length-$u$ cycle in $G_p^{(k)}$. $\square$

**Corollary 12.** *Let $u = \operatorname{lcm}(u_1, u_2, u_3, \ldots, u_n)$, where $u_1, u_2, \ldots, u_n$ are positive integers. If $G_p^{(k)}$ contains a cycle of length $u_i$ for each $i$, then $G_p^{(k)}$ also contains a cycle of length $u$.*

It turns out that for even $k$, the opposite direction is not always true. Later we present a digraph $G_p^{(k)}$ that has a 12-cycle and a 1-cycle but no cycles of length 2, 3, 4, or 6. The following result indicates that this is hardly an isolated occurrence.

**Theorem 13.** *Let $u$ be a composite number and let $k$ be even.*

(1) *If $k \neq 2$ or $u \neq 6$, then there exist infinitely many primes $p$ such that in $G_p^{(k)}$ there exists a length-u cycle but no length-u′ cycles in which $u' \geq 2$ is a positive divisor of $u$.*

(2) *For the case $k = 2$ and $u = 6$, suppose for some prime $p$ that $G_p^{(2)}$ has a cycle of length 6. Then there must also exist a cycle of either length 2 or 3 in $G_p^{(2)}$; furthermore, if $G_p^{(2)}$ has cycles of length 6 and 3 then it must also have a cycle of length 2. The smallest prime $p$ such that $G_p^{(2)}$ has both a length-6 and a length-2 cycle is $p = 19$; in this case, though, $G_{19}^{(2)}$ does not have a length-3 cycle. The smallest prime $p$ such that $G_p^{(2)}$ has cycles of lengths 2, 3, and 6 is $p = 43$.*

Before we start our proof, we need to introduce a very useful lemma proved independently by Bang [1886] and Zsigmondy [1892], as seen in a recent paper by Roitman [1997]:

**Lemma 14** (Bang and Zsigmondy). *Let $k$ and $u$ be integers greater than 1. There exists a prime divisor $q$ of $k^u - 1$ such that $q$ does not divide $k^j - 1$ for all $j$ where $0 < j < u$, except exactly in the following cases:*

(1) *$k = 2^s - 1$, where $s \geq 2$, and $u = 2$;*

(2) *$k = 2$ and $u = 6$.*

*Proof of Theorem 13.* First, let us discuss the case where $k = 2$ and $u = 6$; that is, we suppose there exists a length-6 cycle in $G_p^{(2)}$. By Theorem 4, we must have $\gcd(2^6 - 1, p - 1) > 1$. Now since $2^6 - 1 = 63$, we know $p - 1$ must be divisible by either 7 or 3. Since $2^3 - 1$ is 7 and of course $2^2 - 1$ is 3, we conclude that either $\gcd(2^3 - 1, p - 1) > 1$ or $\gcd(2^2 - 1, p - 1) > 1$ and hence (again by Theorem 4) we must have a cycle of length 3 or length 2. Now suppose (for the sake of argument) that $G_p^{(2)}$ happens to have both a length-6 cycle and a length-3 cycle but no length-2 cycle. If we let $A_i$ represent the number of cycles of length $i$ in the graph of $G_p^{(2)}$, then Theorem 5.6 of [Somer and Křížek 2009] tells us

$$A_6 = \tfrac{1}{6}\big(\gcd(p - 1, 63) - A_1 - 2A_2 - 3A_3\big).$$

Clearly, $A_1 = 1$ since the only nontrivial solution to $x^2 \equiv x \bmod p$ is $x = 1$. We are assuming that $A_2 = 0$ and that $A_3$ and $A_6$ are both positive, and so the above equation becomes

$$A_6 = \tfrac{1}{6}\big(\gcd(p - 1, 63) - 1 - 3A_3\big).$$
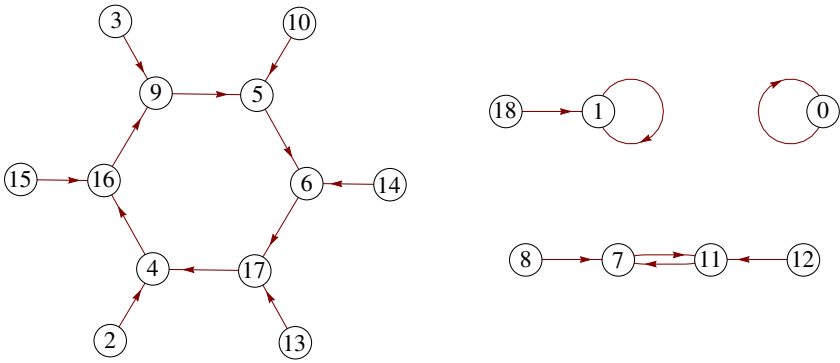
**Figure 1.** The digraph $G_{19}^{(2)}$ has a 6-cycle and 2-cycle but no 3-cycle.

Since $A_3 > 0$, for $A_6$ to be a nonzero integer we must have $\gcd(p-1, 63)$ be either 9, 21, or 63, which are all equivalent to 3 mod 6. But $1 + 3A_3$ will be equivalent to 1 or 4 mod 6, and the difference of these two expressions can never be 0 mod 6, which contradicts $A_6$ being an integer. Hence, the presence of a length-6 cycle and a length-3 cycle really does force there to be a length-2 cycle.

By inspection, $p = 19$ is the smallest prime $p$ such that $G_p^{(2)}$ has a 6-cycle and a 2-cycle; it is easily seen that it does not have a 3-cycle. Also by inspection, $p = 43$ is the smallest prime $p$ such that $G_p^{(2)}$ has a 6-cycle, a 2-cycle, and a 3-cycle. See Figure 1 for the graph of $G_{19}^{(2)}$.

Now if $k \neq 2$ or $u \neq 6$, then in order to prove the theorem it is sufficient to show that there are infinitely many primes $p$ such that for the graph $G_p^{(k)}$ the following conditions hold: for $u_1, u_2, \ldots$ nontrivial proper divisors of $u$,

$$\gcd(k^u - 1, p-1) > 1, \quad \gcd(k^{u_1} - 1, p-1) = 1, \quad \gcd(k^{u_2} - 1, p-1) = 1, \quad \ldots.$$

(By Corollaries 2 and 3, these equations will also imply that the only cycles of length 1 in $G_p^{(k)}$ will be the 1-cycle with vertex 0 and the 1-cycle with vertex 1.) By Lemma 14, we know that there exists a prime divisor $q \mid k^u - 1$ such that $q \nmid k^j - 1$ for $0 < j < u$. Now, consider the set of equivalence relations

$$p - 1 \equiv 0 \bmod q, \tag{1}$$

$$p - 1 \equiv 1 \bmod s, \tag{2}$$

where $s = \operatorname{lcm}(k^{u_1} - 1, k^{u_2} - 1, \ldots)$. Since $q$ is prime, it is obvious that $q \nmid s$ and therefore we can apply the Chinese reminder theorem to get

$$p - 1 \equiv q[q^{-1}]_s \bmod qs,$$

where $[q^{-1}]_s$ is the unique positive integer less than $s$ that is the inverse of $q$ modulo $s$.
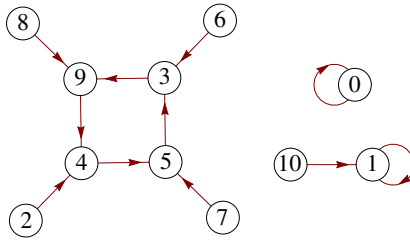
**Figure 2.** The digraph $G_{11}^{(2)}$ has a 4-cycle but no 2-cycle.

Thus, $p = (q[q^{-1}]_s + 1) + qs \cdot n$, where $n \in \mathbb{N}$. Since $q[q^{-1}]_s \equiv 1 \mod s$, we know $q[q^{-1}]_s + 1 \equiv 2 \mod s$, so $\gcd(q[q^{-1}]_s + 1, s) \leq 2$. Now $k$ is even, so $s$ is odd, and we know $\gcd(q[q^{-1}]_s + 1, s) = 1$. On the other hand, it is obvious that $q \nmid q[q^{-1}]_s + 1$; therefore $\gcd(q[q^{-1}]_s + 1, qs) = 1$. Thus, by Dirichlet's theorem, there are infinitely many primes $p$ of the form $p = (q[q^{-1}]_s + 1) + qs \cdot n$, as desired.   □

Now, let us do some examples to illustrate the methods given above.

**Example.** For $p$ a prime, $p \equiv 11 \mod 15$, we know $G_p^{(2)}$ always has a 4-cycle but never has a 2-cycle. This can be shown via the methods of the above proof with $k = 2$, $u = 4$, and $u_1 = 2$, so $q = 5$ and $s = 3$. The two smallest primes $p$ of this type are 11 and 41. The digraph for $G_{11}^{(2)}$ is given in Figure 2.

**Example.** Likewise, using $k = 2$, $u = 9$, and $u_1 = 3$, we can show that for $p$ a prime equivalent to 366 mod 511, the graph $G_p^{(2)}$ always has a 9-cycle but never has a 3-cycle. The smallest prime equivalent to 366 mod 511 is 877, and a partial digraph for $G_{877}^{(2)}$ is given in Figure 3.

**Example.** Finally, using $k = 2$, $u = 12$, and $\{u_1, u_2, u_3, u_4\}$ equal to $\{2, 3, 4, 6\}$, the techniques of our proof of Theorem 13 show that for $p$ a prime, $p \equiv 1262 \mod 4095$, the graph $G_p^{(2)}$ always has a 12-cycle but never has cycles of lengths 2, 3, 4, or 6. The smallest $p$ in this equivalence class is 21737. (This is the smallest prime that arises from the technique of Theorem 13, but it is not the smallest prime $p$ such that $G_p^{(2)}$ has a cycle of length 12 but none of lengths 2, 3, 4, or 6; experimentation shows that the first such prime would be 53, not 21737. We will explain this further in a moment.)

One problem with the above examples (all of which arise from the techniques of Theorem 13) is that while they guarantee an infinite list of primes that satisfy the given requirements, it is not necessarily a complete list. For example, suppose we want to find all primes $p$ such that for $G_p^{(2)}$ we have cycles of length 12, but no cycles of lengths 2, 3, 4, or 6. By our first example, we know any prime of the form $p \equiv 1262 \mod 4095$ will certainly work (and the first prime in this list is 21737). But as mentioned above, $p = 53$ works just fine as well. Let us see if
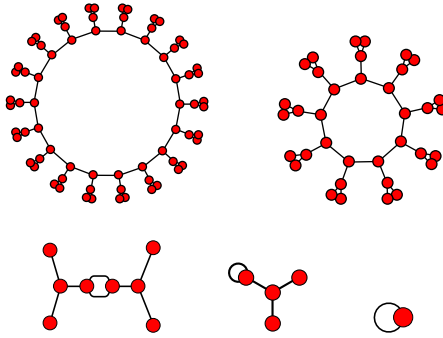
**Figure 3.** The digraph $G_{877}^{(2)}$ has eight components each of the forms shown in the top row, and just one component each of the forms shown in the bottom row. In particular, it has a 9-cycle but no 3-cycle.

we can demonstrate how to find all such primes $p$ such that the digraphs $G_p^{(2)}$ will have cycles of length 12 but not length 2, 3, 4, or 6.

In order to find such $p$'s, we need

$$\gcd(2^{12} - 1, p - 1) > 1, \tag{3}$$

$$\gcd(2^6 - 1, p - 1) = 1, \tag{4}$$

$$\gcd(2^4 - 1, p - 1) = 1, \tag{5}$$

$$\gcd(2^3 - 1, p - 1) = 1, \tag{6}$$

$$\gcd(2^2 - 1, p - 1) = 1, \tag{7}$$

$$\gcd(2^1 - 1, p - 1) = 1. \tag{8}$$

By Lemma 14, we know there exists a prime divisor $q$ such that $q \mid 2^{12} - 1$ but $q \nmid 2^i - 1$, where $i \in \{2, 3, 4, 6\}$, and a brief calculation shows $q = 13$. Therefore, we have $13 \mid \gcd(2^{12} - 1, p - 1)$ and so it follows that $p - 1 = 13n$, or $p = 13n + 1$. In order for (4)–(7) to hold, we must make sure that $n$ does not contain any proper divisor of $(2^i - 1)$, where $i \in \{2, 3, 4, 6\}$; that is, 3, 5, and 7 must not divide $n$. So, a complete list of primes $p$ can be written as the set

$$\{p \text{ is prime} : p = 13n + 1 \text{ where } n \in \mathbb{N} \text{ and } 3, 5, 7 \nmid n\}.$$

The smallest $p$ is indeed 53.

If we glance at (3)–(8), we might wonder if we can modify them to give us more liberty in deciding what cycles we want to have and not have in our $G_p^{(k)}$. Suppose we want to change the above example to have a digraph with cycles of lengths 12 and 2, but no other cycles of lengths 3, 4, or 6. By Theorem 4, we need to start with

$$\gcd(2^{12} - 1, p - 1) > \gcd(2^2 - 1, p - 1) > 1.$$

Since $2^2 - 1 \mid 2^4 - 1$ and $2^2 - 1 \mid 2^6 - 1$, we know that $\gcd(2^4 - 1, p - 1)$ and $\gcd(2^6 - 1, p - 1)$ will both be at least as large as $\gcd(2^2 - 1, p - 1)$, but to prevent $G_p^{(2)}$ from having cycles of length 4 or length 6, we need them to be no larger than $\gcd(2^2 - 1, p - 1)$. Finally, to avoid any 3-cycles, we would like $\gcd(2^3 - 1, p - 1) = 1$. We can satisfy all these requirements if we are able to establish the six equations

$$\gcd(2^{12} - 1, p - 1) = q_2 q_{12}, \tag{9}$$

$$\gcd(2^6 - 1, p - 1) = q_2, \tag{10}$$

$$\gcd(2^4 - 1, p - 1) = q_2, \tag{11}$$

$$\gcd(2^2 - 1, p - 1) = q_2, \tag{12}$$

$$\gcd(2^3 - 1, p - 1) = 1, \tag{13}$$

$$\gcd(2^1 - 1, p - 1) = 1, \tag{14}$$

where $q_2$ and $q_{12}$ are both primes. (Note the similarity between these six equations and the ones given earlier in (3)–(8).)

Fortunately, this is indeed possible. Lemma 14 guarantees that we can find appropriate primes $q_2$ and $q_{12}$; our choices here will be $q_2 = 3$ and $q_{12} = 13$. We also need to ensure that $p - 1$ does not contain any other primes that might also appear in $2^k - 1$ as $k$ runs over the divisors of 12. This can be satisfied by restricting ourselves to the set

$$\{p \text{ is prime} : p = 39n + 1 \text{ where } n \in \mathbb{N} \text{ and } 3, 5, 7 \nmid n\}.$$

It turns out the smallest such $p$ is 79.

Naturally, we seek to generalize this technique, and the following theorem gives the appropriate conditions in which this can be done.

**Theorem 15.** *Let $u \geq 4$ be any composite number, let $k \geq 2$, and let $u' \geq 2$ be a proper divisor of $u$. So long as we do not have either $k = 2$ and $u' = 6$, or $k = 2$ and $u = 6$ and $u' = 3$, then there exist infinitely many primes $p$ such that $G_p^{(k)}$ has both a $u$-cycle and a $u'$-cycle but has no $w$-cycle, where $w$ is any other nontrivial proper divisor of $u$.*

**Remark.** The two restrictions in the above theorem are thanks to Theorem 13, which tells us that every digraph $G_p^{(2)}$ which contains a 6-cycle will also contain either a 2-cycle or 3-cycle, and that if it contains a 6-cycle and 3-cycle then it must also have a 2-cycle.

*Proof of Theorem 15.* We begin by considering the case where $k$ is even and different from 2. This avoids the two exceptions to Lemma 14, and so we know there exist two separate primes $q$ and $q'$ such that $q \mid k^u - 1$ but $q \nmid k^w - 1$ for all $w < u$, and $q' \mid k^{u'} - 1$ but $q' \nmid k^w - 1$ for all $w < u'$. Since $k$ is even, the primes $q$ and $q'$ are

necessarily odd. We want to set up a system similar to the ones in (9)–(14); in this context, our system will be

$$\gcd(k^u - 1, p - 1) = qq', \tag{15}$$

$$\gcd(k^y - 1, p - 1) = q' \qquad \text{if } y < u \text{ and } y \text{ divisible by } u', \tag{16}$$

$$\gcd(k^z - 1, p - 1) = 1 \qquad \text{if } z < u \text{ and } z \text{ not divisible by } u'. \tag{17}$$

(Here, $y$ and $z$ run over the proper divisors of $u$.) These three conditions, along with Theorem 4, would guarantee the existence of a cycle of length $u$ and one of length $u'$ and would prohibit any cycles of length $w$ for $w$ any other nontrivial divisor of $u$. It remains to show there are infinitely many such primes $p$ that satisfy (15)–(17). Fortunately, this is not too hard. Let $Q$ be the product of all the primes other than $q$ and $q'$ that divide $k^u - 1$. Since neither $q$ nor $q'$ divide $k^1 - 1$ and since $k > 2$, there is at least one such prime, and since $k$ is even, all such primes in $Q$ are odd primes. If we now require

$$p - 1 \equiv qq' \bmod (qq')^2, \tag{18}$$

$$p - 1 \equiv 1 \bmod Q, \tag{19}$$

then we are guaranteed (15)–(17), as we now briefly demonstrate.

• To begin with, (18) tells us that $qq'$ divides into $p - 1$, but no higher power of $q$ or $q'$ does so. Also, (19) tells us that no other prime $\rho$ that divides into $Q$ will also divide into $p - 1$. Hence, the gcd's in (15)–(17) must be either 1, $q$, $q'$, or $qq'$.

• To establish (15), we note that by definition, both $q$ and $q'$ divide into $k^u - 1$.

• For (16), we note that $q'$ divides into $k^{u'} - 1$, which divides into $k^y - 1$ for $y$ divisible by $u'$, and that $q$ does not divide into any $k^y - 1$ for $y < u$.

• As for (17), note that $k^z - 1$ is not divisible by $q$ for any $z < u$. If $k^z - 1$ was divisible by $q'$, then $q'$ would divide the gcd of $k^z - 1$ and $k^{u'} - 1$. This gcd is $k^d - 1$ where $d = \gcd(z, u')$ and since $z$ is not divisible by $u'$ then we know $d < u'$, but this contradicts our definition of $q'$. Hence, $k^z - 1$ is not divisible by either $q$ or $q'$ and so we have established (17).

We can now apply the Chinese remainder theorem to write (18) and (19) as $p - 1 \equiv A \bmod Q(qq')^2$ for some integer $A$, which implies

$$p \equiv 1 + A \ \bmod Q(qq')^2$$

Are we now able to apply Dirichlet's theorem to claim that there are infinitely many primes that satisfy the above equivalence? Almost! We need only ensure that $1 + A$ is relatively prime to $Q(qq')^2$. Since (18) tells us that $A \equiv 0 \bmod q$, then $A + 1 \equiv 1 \bmod q$, and the same holds for $q'$. Hence, $A + 1$ is relatively prime to $q$ and to $q'$. Now let $\rho$ be one of the primes that divides $Q$. We know from (19) that $A \equiv 1 \bmod \rho$,

which means $A+1 \equiv 2$ mod $\rho$, but of course $\rho$ is an odd prime, so $A+1$ is relatively prime to $\rho$. We conclude that $A+1$ is relatively prime to $Q(qq')^2$, and so we can apply Dirichlet's theorem to complete the proof (for this case where $k$ even and $k > 2$).

Next, we consider $k = 2$, $u = 6$, and $u' = 2$. This is a very specific case, and if we set $p \equiv 19$ mod 63 to be prime, it is easy to verify that all four of the equations

$$\gcd(2^6 - 1, \, p - 1) = 9, \tag{20}$$

$$\gcd(2^2 - 1, \, p - 1) = 3, \tag{21}$$

$$\gcd(2^3 - 1, \, p - 1) = 1, \tag{22}$$

$$\gcd(2^1 - 1, \, p - 1) = 1. \tag{23}$$

are satisfied. Naturally, there are infinitely such primes $p$ (the first one is $p = 19$) and Theorem 4 tells us that $G_p^{(2)}$ will have a 6-cycle and a 2-cycle but never a 3-cycle.

Next, consider $k = 2$ with neither $u$ nor $u'$ equal to 6. Since this avoids the exceptions to Lemma 14, as before we can find the two separate primes $q$ and $q'$ such that $q \mid k^u - 1$ but $q \nmid k^w - 1$ for all $w < u$, and $q' \mid k^{u'} - 1$ but $q' \nmid k^w - 1$ for all $w < u'$. We would like to define $Q$ to be the product of all primes $\rho$ different from $q$ and $q'$ that divide $2^u - 1$, but it is possible that no such primes $\rho$ exist (consider, for example, $q = 73$ a factor of $2^9 - 1$, and $q' = 7$ a factor of $2^3 - 1$: there are no other prime factors of $2^9 - 1$). If this is the case, simply set $Q = 1$ and proceed as before.

Next, consider when $k > 2$ is odd and we do not have $u' = 2$ and $k = 2^s - 1$ with $s \geq 2$. Lemma 14 gives us the primes $q$ and $q'$ as before, and since $q$ and $q'$ do not divide $k^1 - 1$ (by definition), both $q$ and $q'$ are odd primes. However, in our earlier work, (15)–(17) depended on some of the equations $\gcd(k^z - 1, \, p - 1)$ being equal to 1, but now that $k^z - 1$ is even, this is no longer possible. Instead, we will ask that $p \equiv 3$ mod 4 (which will mean that $p - 1$ is divisible by 2 and not 4), and we seek to establish the system

$$\gcd(k^u - 1, \, p - 1) = 2qq', \tag{24}$$

$$\gcd(k^y - 1, \, p - 1) = 2q' \quad \text{if } y < u \text{ and } y \text{ divisible by } u', \tag{25}$$

$$\gcd(k^z - 1, \, p - 1) = 2 \quad \text{if } z < u \text{ and } z \text{ not divisible by } u'. \tag{26}$$

By Theorem 4 this will be sufficient to create our desired digraph $G_p^{(k)}$. But can we find primes $p$ that satisfy (24), (25), and (26)? Of course! Let $Q$ be the product of all the odd primes other than $q$ and $q'$ that divide $k^u - 1$, with the understanding that if no such primes exist then $Q = 1$. If we now require

$$p - 1 \equiv qq' \text{ mod } (qq')^2, \tag{27}$$

$$p - 1 \equiv 1 \text{ mod } Q, \tag{28}$$

$$p - 1 \equiv 2 \text{ mod } 4, \tag{29}$$

then we are guaranteed (24)–(26), as we now briefly demonstrate:

• As seen earlier, (27) tells us that $qq'$ divides into $p - 1$, but no higher power of $q$ or $q'$ does so. Also, (28) tells us that no other prime $\rho$ that divides into $Q$ will also divide into $p - 1$. And, (29) guarantees that $2 \mid (p - 1)$ but 4 does not. These observations, along with $k - 1$ being even, tell us that the gcd's in (24)–(26) must be either 2, $2q$, $2q'$, or $2qq'$.

• To establish (24), we note that both $p - 1$ and $k^u - 1$ are divisible by $q$ and $q'$.

• For (25), we note that $q'$ divides into $k^{u'} - 1$, which divides into $k^y - 1$ for $y$ divisible by $u'$, and that $q$ does not divide into any $k^y - 1$ for $y < u$. This is identical to our proof for (16).

• Likewise, (26) is proved the same way as (17).

As before, we can now use the Chinese remainder theorem to write $p = 1 + A \bmod 4Q(qq')^2$ for some appropriate $A$, and it is easy to show that $1 + A$ and $4Q(qq')^2$ are relatively prime, thus allowing us to finish the proof by using Dirichlet's theorem.

The very last case to consider is when $k = 2^s - 1$ for $s \geq 2$, and $u' = 2$. The issue here is that $k - 1$ and $k^{u'} - 1$ will have exactly the same prime divisors (just to different powers) so we cannot find an appropriate prime $q'$ as we did earlier, where $q'$ was supposed to divide $k^{u'} - 1$ but not $k - 1$. Instead, we have to proceed as follows. First, choose a prime $q$ such that $q \mid k^u - 1$ but $q \nmid k^w - 1$ for all $w < u$. Note that $q$ is necessarily odd. We now seek to establish

$$\gcd(k^u - 1, p - 1) = 4q, \tag{30}$$

$$\gcd(k^y - 1, p - 1) = 4 \qquad \text{if } y < u \text{ and } y \text{ divisible by 2,} \tag{31}$$

$$\gcd(k^z - 1, p - 1) = 2 \qquad \text{if } z < u \text{ and } z \text{ not divisible by 2.} \tag{32}$$

To do this, we let $Q$ be the (possibly empty) product of all the odd primes other than $q$ that divide into $k^u - 1$, and we require

$$p - 1 \equiv q \bmod q^2, \tag{33}$$

$$p - 1 \equiv 1 \bmod Q, \tag{34}$$

$$p - 1 \equiv 4 \bmod 8. \tag{35}$$

Once more, we can easily show that (33)–(35) imply (30)–(32):

• Equations (33)–(35) imply that $p - 1$ is divisible by $q$ but not $q^2$, by 4 but not 8, and by no other prime factor $\rho$ of $k^u - 1$. Keeping in mind that $k$ is odd, we see that the gcd's in (31) and (32) must be either 2 or 4, and in (30) we must have either $2q$ or $4q$.

• To establish that (30) is equal to $4q$ and not $2q$, we note that $k^u - 1$ is divisible by $k^2 - 1$, which (since $k = 2^s - 1$) is divisible by 4.

- For (31), we note again that $k^u - 1$ is divisible by 4.

- Finally, for $z$ odd, $k^z - 1$ factors as $(k-1)(k^{z-1} + k^{z-2} + \cdots + 1)$. The first expression is $k - 1 = 2^s - 2$, which is divisible by 2 but not 4. The second expression is the sum of an odd number of odd terms, and hence odd. Thus, (32) is indeed equal to 2 and not 4.

As before, we summarize (33)–(35) as a single expression $p = 1 + A \mod 8Qq^2$ for some appropriate $A$, and it is now fairly routine to finish the proof by using Dirichlet's theorem. $\qquad\square$

## Acknowledgment

## References

[Bang 1886] A. S. Bang, "Taltheoretiske undersølgelser", *Tidsskrift f. Math.* **4**:5 (1886), 70–80, 130–137.

[Gallian 2010] J. A. Gallian, *Contemporary abstract algebra*, 7th ed., Brooks/Cole, Belmont, CA, 2010.

[Lucheta, Miller and Reiter 1996] C. Lucheta, E. Miller, and C. Reiter, "Digraphs from powers modulo $p$", *Fibonacci Quart.* **34**:3 (1996), 226–239. MR Zbl

[Niven, Zuckerman and Montgomery 1991] I. Niven, H. S. Zuckerman, and H. L. Montgomery, *An introduction to the theory of numbers*, 5th ed., Wiley, New York, 1991. MR Zbl

[Roitman 1997] M. Roitman, "On Zsigmondy primes", *Proc. Amer. Math. Soc.* **125**:7 (1997), 1913–1919. MR Zbl

[Somer and Křížek 2004] L. Somer and M. Křížek, "On a connection of number theory with graph theory", *Czechoslovak Math. J.* **54(129)**:2 (2004), 465–485. MR Zbl

[Somer and Křížek 2009] L. Somer and M. Křížek, "On symmetric digraphs of the congruence $x^k \equiv y \pmod{n}$", *Discrete Math.* **309**:8 (2009), 1999–2009. MR Zbl

[Wilson 1998] B. Wilson, "Power digraphs modulo $n$", *Fibonacci Quart.* **36**:3 (1998), 229–239. MR Zbl

[Zsigmondy 1892] K. Zsigmondy, "Zur Theorie der Potenzreste", *Monatsh. Math. Phys.* **3**:1 (1892), 265–284. MR Zbl

dresdeng@wlu.edu                *Department of Mathematics, Washington and Lee University, Lexington, VA, United States*

wenda-tu@uiowa.edu              *Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA, United States*

# Enumerating spherical $n$-links

## Madeleine Burkhart and Joel Foisy

(Communicated by Jim Hoste)

We investigate spherical links: that is, disjoint embeddings of 1-spheres and 0-spheres in the 2-sphere, where the notion of a split link is analogous to the usual concept. In the quest to enumerate distinct nonsplit $n$-links for arbitrary $n$, we must consider when it is possible for an embedding of circles and an even number of points to form a nonsplit link. The main result is a set of necessary and sufficient conditions for such an embedding. The final section includes tables of the distinct embeddings that yield nonsplit $n$-links for $4 \leq n \leq 8$.

## 1. Introduction

The enumeration of links in 3-space is well-studied [Hoste 2005]. However, there has not been much study of a planar/spherical analog outside the confines of its appearance in graphs [Archdeacon and Sagols 2002]. We aim to get the ball rolling on spherical links.

An $n$-link $\mathcal{L}$ in the 2-sphere is a disjoint collection of $q$ embedded 1-spheres and $n - q$ embedded 0-spheres. Two links are equivalent if there is a spherical isotopy taking one to the other. Throughout this paper we use standard notation for a $k$-sphere: $S^k$. When speaking of spherical links, it does not make topological sense to call 0-spheres "components", since an entire $S^0$ is not connected. Henceforth we will refer to an $S^1$ or an $S^0$ as a *piece* of an $n$-link. We will call a spherical embedding of 1-spheres a *nesting*. Note that when we refer to nestings and nests in this paper, we are working with entities distinct from those in [Archdeacon and Sagols 2002].

We must now consider what constitutes a split spherical link. Note that the following definition only makes sense after we have chosen which pairs of points form 0-spheres: An $n$-link $\mathcal{L}$ is *split* if there exists an embedding $\phi$ of $S^1$ in $S^2 - \mathcal{L}$ such that each component of $S^2 - \phi(S^1)$ contains at least one piece of $\mathcal{L}$ and each piece of $\mathcal{L}$ is entirely contained in one such component. Otherwise, $\mathcal{L}$ is nonsplit.
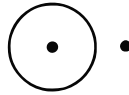
**Figure 1.** A nonsplit 2-link.
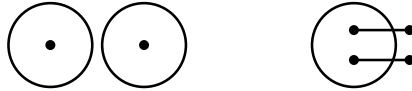


**Figure 2.** A 3-link with two circles and a 3-link with one circle.

Although there is only one type of nonsplit spherical 2-link, when we look at $n$-linking for $n > 2$, we can have different numbers of disjoint 1-spheres and 0-spheres. For example, we could have the two types of nonsplit spherical 3-links as in Figure 2.

We will find that the enumeration of $n$-link-types becomes more richly complex as $n$ increases. Before finding all $n$-links for $n \leq 8$ in Section 3, we lay down the necessary and sufficient conditions for any spherical embedding of $q$ circles and $2\ell$ points to form a nonsplit $(q+\ell)$-link (given appropriate $S^0$ identifications). When considering such links, it will be helpful not only to think about nestings with points, but also to associate a weighted tree $\mathcal{T}$. To construct $\mathcal{T}$, first consider the nesting $\mathcal{N}$. If we identify a vertex on each circle, this embedding is a plane graph of disjoint loops, so the dual graph will be a tree in which each vertex is an open component of $S^2 - \mathcal{N}$. To account for embedded points, we give each vertex a weight equal to the number of points in the corresponding region.

The weighted tree $\mathcal{T}$ corresponds to a nesting with unpaired points, but we want to work with links; we will need to consider what happens to the tree after we make $S^0$ identifications. To make an identification, we will choose two vertices that each have weight at least 1, add an edge between them, and reduce their weights each by one (Figure 3).
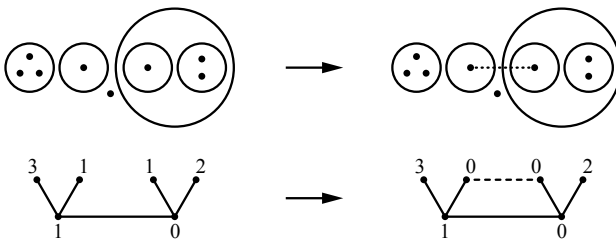


**Figure 3.** A nesting with points and its corresponding tree as we make an $S^0$ identification.
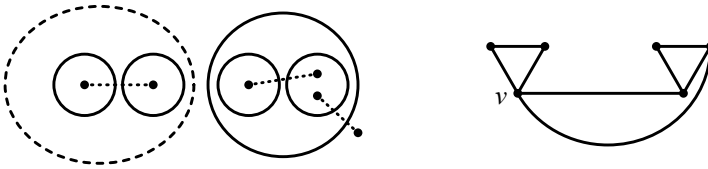
**Figure 4.** The vertex $v$ corresponds with the region of the splitting circle.

If we do this until each vertex has weight 0, the resulting multigraph $G_{\mathcal{T}}$ will represent a link (unique if we distinguish the original tree edges from the $S^0$ identification edges). How can we tell from the graph if the link is split? Certainly a loop in the graph represents a split $S^0$. Any other type of split link, in which both components of $S^2 - \phi(S^1)$ (as in the split definition) have some positive number of circles, occurs if and only if there is a cut vertex in the multigraph (Figure 4).

We have now built up enough background to state our main result in dual ways.

**Theorem 1.1.** *Suppose we have a weighted tree $\mathcal{T}$ with $q$ edges and total weight $2\ell$. In the corresponding embedding of $q$ circles (with nesting $\mathcal{N}$) and $2\ell$ points, it is possible to identify $0$-spheres so that we have a nonsplit spherical $(q+\ell)$-link if and only if all of the following conditions are satisfied:*

(1) *Each leaf has weight at least one. That is, we must embed at least one point in each simply connected region of $S^2 - \mathcal{N}$.*

(2) *No vertex $v$ is assigned a weight greater than $\ell - \deg(v) + 1$. That is, we can embed no more than $\ell - \kappa + 1$ points in a region of $S^2 - \mathcal{N}$ that has fundamental group $\mathbb{Z} * \cdots * \mathbb{Z}$, where $\mathbb{Z}$ appears $\kappa - 1$ times.*

(3) *Given any vertex of degree $\kappa$, the other vertices have total weight summing to at least $2(\kappa - 1)$. In other words, given a region as in (2), we must embed at least $2(\kappa - 1)$ points in the remaining regions.*

With this result, we can tell which embeddings of $n$ (1 and 0)-spheres will form a nonsplit $n$-link. However, enumeration will require distinguishing links from one another on the sphere, which we only address for $n \leq 8$ in this paper.

### Future directions

All the enumeration in this paper was done by hand; code will probably be necessary to enumerate spherical $n$-links for $n \geq 9$. As there is a one-to-one correspondence between nestings and unlabeled trees, much of the code will probably be similar to what is used in the problem of enumerating unlabeled trees (see [Harary 1969; Sloane 2006]).

While our results regard embeddings in $S^2$, it would be interesting to see how tabulations differ on different surfaces; for example, while a spherical embedding

yields a correspondence between nestings and unlabeled trees, in the plane the correspondence is between nestings and rooted trees.

Our necessary and sufficient conditions depend on "appropriate" $S^0$ identifications. What happens if we make the worst possible $S^0$ identifications; that is, given a nesting with an even number of disjointly embedded points, what is the minimal nonsplit $n$-link among all possible $S^0$ pairings?

We could seek to generalize our result in a combinatorial manner; instead of looking at 0-spheres (i.e., pairs of points), we could look at triples, quadruples, or $\lambda$-tuples of points.

Because of the Jordan–Brouwer separation theorem [Guillemin and Pollack 1974], our results generalize to higher dimensions. The same necessary and sufficient conditions and link enumerations apply to embeddings of $k$-spheres and 0-spheres in $S^{k+1}$, since the dual weighted tree construction will still be well-defined. Perhaps this result has applications. It would also be interesting to investigate enumerating other types of higher-dimensional linking with spheres of different dimensions.

## 2. Proof of Theorem 1.1

In the following lemmas, we will switch between thinking about nestings and weighted trees. The following concepts will be useful when working with nestings.

Suppose we have a nesting $\mathcal{N}$. If we single out an open region in $S^2 - \mathcal{N}$ there will be some number of embedded circles that form holes in the region. We will call each such circle, along with all pieces in its interior, a *nest* (see Figure 5).

Suppose we have a nesting $\mathcal{N}$ and single out an open region $R$. Let each nest relative to $R$ have a corresponding vertex. We add an edge between vertices if there is an $S^0$ identification "connecting" the nests. We will denote any graph resulting from this process as $H_R$.

**Lemma 2.1.** *The conditions of Theorem 1.1 are necessary for a $(q+\ell)$-link.*

*Proof.* Condition (1) is obvious; if a leaf $v$ has weight 0, no matter how we construct $G_{\mathcal{T}}$ from $\mathcal{T}$, the vertex $v$ will still have degree 1 and so the resulting graph cannot be 2-connected. Now suppose we single out a region $R$. The number $\kappa$ of nests relative to $R$ is equal to the number of vertices in $H_R$. To ensure a nonsplit link, we must make $S^0$ identifications so that $H_R$ is connected; minimally, we will thus
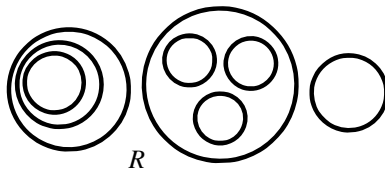


**Figure 5.** There are three nests relative to the region $R$.
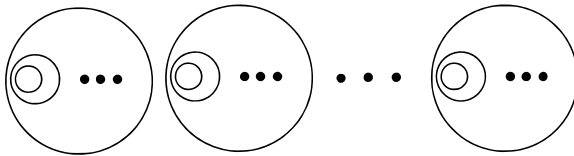
**Figure 6.** A simple $\kappa$-nesting.

need $(\kappa - 1)$ 0-spheres and thus $2(\kappa - 1)$ points in the non-$R$ regions. This proves the necessity of (3). Condition (2) follows: since we need $2(\kappa - 1)$ points among the non-$R$ regions to connect them and since we are avoiding split 0-spheres, $R$ cannot have over half of the remaining $2\ell - 2(\kappa - 1)$ points. $\qquad\square$

We now prove sufficiency in a specific base case before proving it in general. In this proof we will primarily refer to nestings rather than weighted trees. We define a *simple $\kappa$-nesting* to be an embedding of $q$ 1-spheres in $S^2$ that can achieve the arrangement of simple nests as in Figure 6 through spherical isotopy. The corresponding tree is a (possibly topologically nonreduced) star.

Given a simple nesting $\mathcal{N}$, we will call the $\kappa$ simply connected regions of $S^2 - \mathcal{N}$ *innermost* (corresponding to leaves). We call the region with fundamental group $\mathbb{Z} * \cdots * \mathbb{Z}$ (where $\mathbb{Z}$ appears $\kappa - 1$ times) *outermost*. Any other region (though there need not be any regions beyond the innermost and outermost ones) in a simple $\kappa$-nesting is annular, with fundamental group $\mathbb{Z}$. When we refer to nests in a simple nesting, we will always work relative to the outermost region, denoting $H_R$ as just $H$.

**Lemma 2.2.** *The conditions in Theorem 1.1 are sufficient for a nonsplit $(q+\ell)$-link in a simple $\kappa$-nesting $\mathcal{N}$.*

*Proof.* We will first find a way to link the circles and then the 0-spheres. Given that we use exactly $2(\kappa - 1)$ points in the former and that no innermost or annular region has more than $\ell - \kappa + 1$ (i.e., more than half of the) unpaired points after the process, the latter will follow easily. Because we want $H$ to be connected while only matching $(\kappa - 1)$ 0-spheres, it is imperative to avoid cycles during the construction. We now state our algorithm for linking all $q$ circles and $2(\kappa - 1)$ of the $2\ell$ points given an embedding that follows the conditions of Theorem 1.1:

(1) Pick a region $R$ with the most unpaired points; in case of ties, let $R$ be in a nest $N_R$ with the most total unpaired points. Pair one point from $R$ (the *selector*) with a point (the *selected*) in another nest. If possible, let the selected point come from as-yet unchosen innermost region, making sure such a pairing does not induce a cycle in $H$. If our choice of $R$ leads inevitably to either a cycle or a pairing that does not include an as-yet unchosen innermost region when such a thing exists, we adjust our choice of our selector region $R_1$ to be in a different nest-component (i.e., a collection of nests whose vertices in

$H$ are in a different component from the vertex corresponding to $N_R$). Let $R_1$ have the most unpaired points of the regions in different nest-components from $N_R$, preferably in a nest with the most total points. Then pair a selector point from $R_1$ with a point in an as-yet unchosen innermost region.

(2) Mark off this $S^0$ so the points are disregarded for the rest of the algorithm.

(3) Repeat steps 1–2 until $(\kappa - 2)$ 0-spheres have been paired off.

(4) If each of the $(\kappa - 2)$ 0-spheres contains a point from an annular region, match a point each from the two remaining innermost regions for the last $S^0$. If not, follow steps 1–2 for the last $S^0$.

In this algorithm, we form exactly $(\kappa - 1)$ 0-spheres, so it remains to prove:

(a) that we are indeed allowed to choose points in the first step without inducing cycles given only the conditions in the theorem,

(b) that the algorithm results in a nonsplit $(q+\kappa-1)$-link, and

(c) that no innermost or annular region is left at the end of the construction with more than $\ell - \kappa + 1$ unpaired points.

(a) At some point in the algorithm, let $R_0$ be our initial choice for $R$ in Step (1) and let the vertex $v$ represent $R_0$'s nest in $H$. Suppose we have not yet had to switch $R$. If an innermost region $I$ in a nest whose vertex is disconnected from $v$ does not yet have a matched point, we can match a point from $R_0$ with one from $I$ without inducing a cycle. Now suppose that every innermost region in $\mathcal{N}$ has matched points. Because of the rules for choosing the initial $R$ in each step, every nest-component will have extra points; we can use one such point in a distinct nest-component to pair with one of the $R_0$'s without inducing a cycle.

Now suppose we are in the remaining situation: The nests with vertices in components disconnected from $v$ each have matched innermost regions, but at least one nest (with corresponding vertex $u$) in $v$'s connected component $C$ in $H$ has no matched points in its innermost region. Note that $u$ and $v$ are not necessarily distinct, but we will not have to deal with this contingency until we prove (c).

Consider the nesting corresponding to $C$. Because we still have an unmatched innermost region, all prior matchings had their selected points in distinct innermost regions. Thus, since $C$ is a tree (being connected with no cycles) and the nest corresponding to $u$ has an unmatched innermost region, the rest of the nests corresponding to $C$'s non-$u$ vertices must have matched innermost regions. In fact, since we assumed all the non-$C$ nests had matched innermost regions, the $u$-nest is the only one without a match in all of $\mathcal{N}$. Ergo when we switch $R$, we will not have to do it again for the rest of the construction. Note also that when we switch the selector region, we have only one choice for the region of the selected point: it must be a point in the innermost region of the $u$-nest.
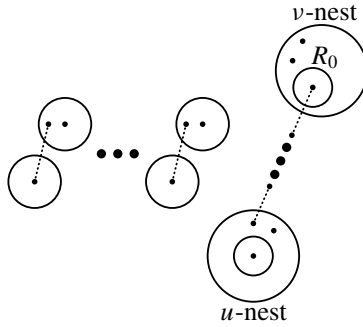
**Figure 7.** A simplified diagram of the situation when we have to switch $R$.

In addition, since the $C$-nesting has extra points and the nest-component containing $R_1$, which may no longer have extra points after deposing $R_0$, is now matched with the $C$-nesting, we can proceed as usual: all the remaining nest-components have extra points. Thus it is possible to follow our construction given the conditions of Theorem 1.1.

(b) The somewhat strict stipulations in the algorithm have a great payoff: since $H$ has $\kappa - 1$ edges and no cycles, it is a tree, and thus connected. Since we also make sure that every innermost region has a point matched with one in another nest, it follows immediately that all the circles are linked nontrivially.

(c) We will consider four cases to prove that no innermost or annular region is left with more than $\ell - \kappa + 1$ unpaired points. Before delving in, however, we make note that since we have already paired $2(\kappa - 1)$ of the $2\ell$ points, we will be left with $2\ell - 2\kappa + 2$ unpaired points; a region is left with more than $\ell - \kappa + 1$ unpaired points after the algorithm if and only if it has two or more unpaired points than any other region in $\mathcal{N}$:

(i) Suppose we finish the algorithm with an innermost region $I$ having more than $\ell - \kappa + 1$ points left over and we never had to switch $R$. Since $I$ ends with at least two more points than any other region, since the algorithm only matches a point at a time from any one region, and since $I$ never had its "$R$" status revoked for the "special case" stipulations in the construction, $I$ must have been $R$ at each step. When we add in the $\kappa - 1$ points we matched in $I$, we find that $I$ must have started with more than $\ell$ points, a contradiction.

(ii) Suppose we finish the algorithm with an annular region $A$ having more than $\ell - \kappa + 1$ points left over and we never had to switch $R$. This case and its corresponding argument are an analog to those of (i) except for the stipulation in Step (4) of the construction; no matter, for when we add the $\kappa - 2$ matched points to $A$'s total, we find that $A$ must have started with more than $\ell - 1$ points, another contradiction.

(iii) Now suppose we have to switch $R$ at some point in the algorithm and, letting $R_0$, $v$, $u$, and $C$ be as above, $v$ is distinct from $u$. Let $m$ be the number of unpaired points in $R_0$ at this step. Since the $u$-nest has an unmatched innermost region, by the rules of the algorithm, it could never have been a selected nest. But since $u$ is connected to $v$, the $u$-nest must have had an annular selector region $A$ that at some prior step in the algorithm had a number of unpaired points greater than or equal to $R_0$'s then-number of unpaired points. It follows that $R_0$ and $A$ (and any other appropriate regions) traded off being $R$ according to the usual rules, implying that $m$ must be no more than one greater than the number of points in $A$.

When we apply the switch, the number of unmatched points in $A$ and in $R_0$ remains the same. If the construction is not yet finished, we can continue in the usual way (the stipulation in Step (4) will not apply since we have already matched all the innermost regions), with $R_0$, $A$, and any other appropriate regions trading off as $R$. However, no matter what, we will not have an end situation in which a region has at least two more points than any other region. Thus, no annular or innermost region is left with more than $\ell - \kappa + 1$ points.

(iv) Lastly, suppose we have to switch $R$ at some point in the algorithm and $v = u$. Let $m$ be as in (iii). We can assume that $m$ is strictly greater than the number of points in any other region; if there were equality, we wouldn't risk ending the algorithm with $m$ having two more points than any other region. Note that there is at least one innermost region $I$ in each nest-component (distinct from the $C$-nesting) that trades off being $R$ with $R_0$ until the $R$ switching step. Hence, $m$ is exactly one greater than the number of points in at least one other region at the $R$ switching step. We can narrow our focus to the case where there is only one component distinct from $C$. If there were not, in the step after switching $R$, a point in $R_0$ would pair with a point in another component (which has a region with $m-1$ points), thus preventing $R_0$ from finishing the algorithm with two more points than any other region.

In the case of only two components, the $R$ switching step is the last step of the construction. Thus, we must show that $m \le \ell - \kappa + 1$. Consider the situation at the beginning of the $R$ switching step. Because the $v$-nest still has an unmatched innermost region, it has strictly more than $m$ points. Thus, by the rules of choosing $R$ in case of ties, the nest containing $I$ must have some annular region with $a$ points, where $a \ge 1$. Figure 8 illustrates the situation. At this stage we need at least $m + (m - 1) + a + 1 + 2(\kappa - 2) = 2m + a + 2(\kappa - 2)$ of the $2\ell$ total points. Now suppose that $m \ge \ell - \kappa + 2$. Then we have at least $2(\ell - \kappa + 2) + a + 2(\kappa - 2) = 2\ell + a$ points. But since $a \ge 1$, we have reached a contradiction. Thus no region in this case is left with more than $\ell - \kappa + 1$ points after the construction.

It now only remains to show that we can pair up the unmatched points so that there are no split 0-spheres. To do so, we use the following algorithm:
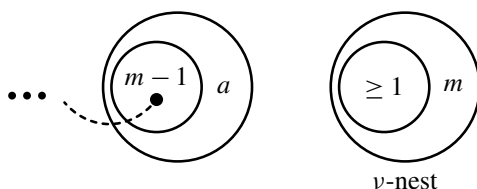
**Figure 8.** The situation before switching $R$ when $v = u$.

(1) Pick two regions $A$ and $B$, each having a number of points greater than or equal to that of any other region in $S^2 - \mathcal{N}$.

(2) Form an $S^0$ from a point in $A$ and a point in $B$.

(3) If there are still unpaired points, return to the first step. If not, we are done.

Suppose that we have followed through with this algorithm but still have at least one split $S^0$ in some region $\mathcal{R}$. Note that $\mathcal{R}$ is the only region left with unpaired points; if there were others we could continue with the algorithm. In addition, if we run the algorithm backwards, one of the two most recently matched points must have come from $\mathcal{R}$. In fact, this is true at each step: since $\mathcal{R}$ has at least two more points than any other region at the end of each step, it must have been one of the regions with the most points at the beginning of any step. Thus, if we run the algorithm all the way back ($\ell - \kappa + 1$ steps), counting the number of points in $\mathcal{R}$ along the way, we find that $\mathcal{R}$ must have started this second algorithm with at least $\ell - \kappa + 2$ points, a contradiction.

Thus it is possible to find a nonsplit $(\ell+q)$-link in a simple $\kappa$-nesting given the conditions of Theorem 1.1. $\qquad\square$

We can now show sufficiency for any nesting with points.

**Lemma 2.3.** *The conditions of Theorem 1.1 are sufficient for a $(q+\ell)$-link*

*Proof.* We will use induction on the number of vertices in the weighted tree. Lemma 2.2 covered the base case, so assume that the conditions of the theorem are sufficient for a $(q-1+\ell)$-link (i.e., on any tree with $q$ vertices). Let $\mathcal{T}_0$ be a weighted tree with $q + 1$ vertices that follows the conditions of Theorem 1.1.

Let $v_0$ be a leaf of $\mathcal{T}_0$ with weight $\mu_1$ and let $u_0$ be the vertex adjacent to $v_0$, with weight $\mu_2$. Now suppose we delete $v_0$ from $\mathcal{T}_0$ and absorb its weight into $u_0$. From this we get a new weighted tree $\mathcal{T}_1$, where $u_1 \in V(\mathcal{T}_1)$ used to be $u_0$. Note that $\deg(u_1) = \deg(u_0) - 1$. Obviously this move preserves the first condition of Theorem 1.1 in $\mathcal{T}_1$. Suppose first that the move preserves the second condition: $u_1$'s weight, $\mu_1 + \mu_2$, is less than or equal to $\ell - \deg(u_1) + 1 = \ell - \deg(u_0) + 2$.

We first aim to show that $\mathcal{T}_1$ follows the third condition given that it follows the second. Let $\mu_3 = 2\ell - (\mu_1 + \mu_2)$. We want to show that $\mu_3 \geq 2(\deg(u_1) - 1) =$

$2(\deg(u_0) - 2)$. Because $\mathcal{T}_0$ follows the rules, we have

$$\mu_1 + \mu_3 \geq 2(\deg(u_0) - 1), \tag{1}$$

and because of our assumption on $\mathcal{T}_1$,

$$\mu_1 + \mu_2 \leq \ell - \deg(u_0) + 2. \tag{2}$$

By the bound given by (2) and the definition of $\mu_3$, we have $\mu_3 \geq \ell + \deg(u_0) - 2$, so if $\ell \geq \deg(u_0) - 2$, we're in the clear. Henceforth assume that $\ell \leq \deg(u_0) - 3$. From (1), we have $\mu_3 \geq 2(\deg(u_0) - 1) - \mu_1$. Using the upper bound on $\mu_1$ from (2) and the one on $\ell$, we obtain

$$\begin{aligned}
\mu_3 &\geq 2(\deg(u_0) - 1) - (\ell - \deg(u_0) + 2 - \mu_2) \\
&= 2(\deg(u_0) - 2) - \ell + \deg(u_0) + \mu_2 \\
&\geq 2(\deg(u_0) - 2) - (\deg(u_0) - 3) + \deg(u_0) + \mu_2 \\
&= 2(\deg(u_0) - 2) + 3 + \mu_2 \\
&> 2(\deg(u_0) - 2),
\end{aligned}$$

which we sought.

We have thus shown that if we choose a $v_0$ to delete such that $\mathcal{T}_1$ follows the second condition, it will follow all the conditions. Thus, we can use the inductive assumption to add edges to $\mathcal{T}_1$ using the weights so that the resulting multigraph $G_{\mathcal{T}_1}$ is 2-connected and contains no loops. When we add $v_0$ back (along with edge $v_0 u_0$), we transfer $\mu_1$ of $u_1$'s added edges to $v_0$. This operation will certainly not create any loops. We now show that it preserves 2-connectivity. Consider any vertex $w$ that is not $v_0$ or $u_0$: the operation preserves the two internally disjoint paths between any two vertices that are not $v_0$ or $u_0$, so if we were to delete $w$, all the other non-($u_0$ or $v_0$) vertices would remain connected. But $u_0$ and $v_0$ would also be connected to the rest of the graph since they are connected to each other and at least one other non-$w$ vertex. Now suppose we delete $u_0$ from $\mathcal{T}_0$. Again, all the non-$v_0$ vertices will still be connected. But $v_0$ will also be connected to the rest of the graph since it is adjacent to at least one non-$u_0$ vertex. Lastly, suppose we delete $v_0$: the rest of the graph is still connected by the $\mathcal{T}_1$ edges. Thus, the multigraph $G_{\mathcal{T}_0}$ induced by $G_{\mathcal{T}_1}$ is 2-connected and without loops and thus determines a nonsplit planar $(q+\ell)$-link

It now only remains to show that we can pick a $v_0$ to remove such that $u_1$ has weight less than or equal to $\ell - \deg(u_1) + 1$. Suppose we cannot find such a $v_0$. Let $\lambda$ be the number of leaves in $\mathcal{T}_0$ and let $\kappa = \max\{\deg(v) : v \in V(\mathcal{T}_0)\}$. Since we have already shown the result for simple nestings in Lemma 2.2, we can assume $\kappa \leq \lambda - 1$, that $\lambda \geq 4$, and that there are at least two $u_0$s we could have depending on our choice of $v_0$. Also, since any $u_0$ has weight less than or equal to $\ell - \deg(u_0) + 1$ and the corresponding $u_1$ has weight greater than or equal to $\ell - \deg(u_0) + 3$, each leaf must

have weight at least 2. Thus the total weight of $\mathcal{T}_0$ is at least $2(\ell-\kappa+3)+2(\lambda-2) \geq 2\ell-2\kappa+6+2(\kappa-1)=2\ell+4$, a contradiction. Thus we are able to choose a "nice" $v_0$ such that the inductive hypothesis holds and is inherited by the larger tree. $\qquad\square$

## 3. Enumeration

We mentioned in the Introduction that there is only one nonsplit spherical 2-link and there are two types of nonsplit spherical 3-links. We have now proven which embeddings will form nonsplit links given appropriate $S^0$ identifications. However, enumeration encompasses even more complications: we must determine whether an embedding is unique up to spherical isotopy. In addition, we have a couple different ways to count links: we can simply count the allowable embeddings or we can count how many ways we can identify 0-spheres appropriately within an embedding (Figure 9). In the link diagrams found in the online supplement, if there is more than one allowable $S^0$ identification for an embedding, we will write how many total identifications there are next to its image. Note that there are four distinct nonsplit 4-links; 11 distinct embeddings and 12 distinct 5-links; 32 distinct embeddings and 39 total 6-links; 105 total embeddings and 158 total 7-links; and 354 embeddings and 723 8-links.

To show rigorously how many allowable $S^0$ pairings there are in an embedding, one fact is particularly helpful: The number of $S^0$ identifications between two regions $R_1$ and $R_2$ is greater than or equal to $p(R_1)+p(R_2)-m$, where $p(R)$ is the number of unmatched points in $R$ and $m$ is half the number of unmatched points left in a nesting. This fact is easily proven. Let $r=p(R_1)+p(R_2)$. If $r \leq m$, the claim is trivially true. Suppose $r>m$. Then there are not enough points in the rest of the nesting to fully match with points in $R_1$ and $R_2$ without inducing split 0-spheres; we must match a point in $R_1$ with one in $R_2$. Now $r$ has decreased by two and $m$ has decreased by one. If $r-2>m-1$, we again match a point from $R_1$ with one from $R_2$. We must iterate $k$ times, where $r-2(k-1)>m-(k-1)$ and $r-2k \leq m-k$: that is when $k=r-m$, which we sought.

The above allows us to reduce larger cases to smaller cases. We also use commonsense techniques, such as choosing a region that can be distinguished from others (usually one with one point) to determine exhaustively the matching possibilities or utilizing symmetry without loss of generality.
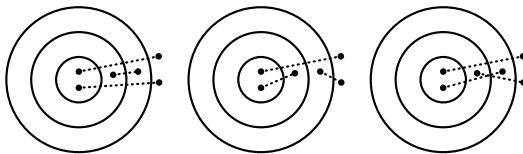


**Figure 9.** This is the same embedding, but there are three different 6-links here.

## Acknowledgements

## References

[Archdeacon and Sagols 2002]  D. Archdeacon and F. Sagols, "Nesting points in the sphere", *Discrete Math.* **244**:1-3 (2002), 5–16. MR Zbl

[Guillemin and Pollack 1974]  V. Guillemin and A. Pollack, *Differential topology*, Printice-Hall, Englewood Cliffs, NJ, 1974. MR Zbl

[Harary 1969]  F. Harary, *Graph theory*, Addison-Wesley, Reading, MA, 1969. MR Zbl

[Hoste 2005]  J. Hoste, "The enumeration and classification of knots and links", pp. 209–232 in *Handbook of knot theory*, edited by W. Menasco and M. Thistlethwaite, Elsevier, Amsterdam, 2005. MR Zbl

[Sloane 2006]  N. J. A. Sloane, "Number of trees with $n$ unlabeled nodes", 2006, available at http://oeis.org/A000055.

burkhm2@uw.edu            *Mathematics Department, University of Washington, Seattle, WA, United States*

foisyjs@potsdam.edu       *Department of Mathematics, SUNY Potsdam, Potsdam, NY, United States*

# Double bubbles in hyperbolic surfaces

Wyatt Boyer, Bryan Brown, Alyssa Loving and Sarah Tammen

(Communicated by Michael Dorff)

We seek the least-perimeter way to enclose and separate two prescribed areas in certain hyperbolic surfaces.

## 1. Introduction

The isoperimetric problem of enclosing a given area in a least-perimeter way has been investigated in various surfaces. The classical isoperimetric theorem in the plane asserts that the circle is the shortest curve to enclose a given area in the plane. While this result is widely known, the solution of the isoperimetric problem has proved to be elusive in surfaces aside from the plane. By 1999, the problem had been solved for a handful of Riemannian surfaces, namely, the Euclidean plane, a round sphere, a round projective plane, the hyperbolic plane, a circular cone, a circular cylinder, a flat torus or Klein bottle, and a general surface of revolution [Howards et al. 1999]. Adams and Morgan [1999] obtained further results in hyperbolic surfaces. The related problem of discovering the least perimeter needed to enclose and separate two given volumes has invited exploration as well.

Particular interest has been garnered by the double bubble conjecture. The double bubble conjecture states that three spherical caps meeting at $\frac{2\pi}{3}$ angles (the "standard double bubble") is the least-perimeter way to enclose and separate two given volumes. This has been believed to be true since the nineteenth century, but it was first articulated as a conjecture by Joel Foisy [1991], an undergraduate student at Williams College, in his senior thesis, and it was proved in the planar case in [Foisy et al. 1993]. Joel Hass, Michael Hutchings, and Roger Schlafly [Hass et al. 1995] attacked the conjecture in the $\mathbb{R}^3$ case using heavily computational methods, successfully resolving the problem for the case where the two volumes are equal. Finally, Michael Hutchings, Frank Morgan, Manuel Ritoré, and Antonio Ros [Hutchings et al. 2002] proved the double bubble conjecture for any ratio of two volumes in $\mathbb{R}^3$. Moreover, Andrew Cotton and David Freeman [2002] have

shown the conjecture to hold for the hyperbolic plane as well as the case of equal volumes in hyperbolic 3-space. In certain hyperbolic surfaces however, the standard double bubble is not perimeter-minimizing. We study this problem, following the work on single bubbles by Adams and Morgan [1999].

Section 2 discusses the existence and regularity of perimeter-minimizing double bubbles. Section 3 considers $n$-punctured spheres. Proposition 3.6 identifies small perimeter-minimizing double bubbles as horocycles around cusps. Section 4 focuses on double bubbles on the thrice-punctured sphere. Conjecture 4.1 describes perimeter-minimizing double bubbles as horocycles for small areas and $\theta$-curves for large areas. Proposition 4.2 shows that, for equal areas, $\theta$-curves are shorter than horocycles for a specific range of areas through direct computations. Propositions 4.7–4.9 show necessary conditions on the topology of perimeter-minimizing double bubbles using inequalities obtained in Lemmas 4.3–4.5. Section 5 considers the once-punctured torus. Proposition 5.1 proves that for relatively small areas two horocycles around a cusp are shorter than a horocycle with a lens.

## 2. Existence and regularity

**Definition 2.1.** A *double bubble* on a surface consists of two disjoint open regions with piecewise smooth boundaries. The *perimeter* refers to the union of the boundaries or its length. We do not assume that each region, or that the perimeter, or that the entire bubble (the union of the regions and the perimeter) is connected. We call the bubble perimeter-minimizing or sometimes just minimizing if it minimizes perimeter for fixed area of each region.

Morgan [1994] examined existence and regularity for soap bubble clusters in $\mathbb{R}^2$ and on compact Riemannian surfaces, and his results and proofs apply to geometrically finite hyperbolic surfaces.

**Theorem 2.2** (existence and regularity). *In a complete hyperbolic surface, there exists a least-perimeter double bubble, enclosing and separating two regions of prescribed areas. Its perimeter consists of curves of constant curvature meeting in threes at angles of $\frac{2\pi}{3}$; all curves separating a specific pair of regions have the same curvature.*

*Proof.* We explain the extension of Morgan [1994] to the noncompact case. If in a minimizing sequence a region goes out a cusp, its area goes to 0 and it may be discarded. If it goes out a flared end, it can be translated back inside a compact region.                                                                                  □

We are assuming that the sum of the two areas is less than the area of the surface; the complement is a third region. It remains conjectural in general that each of the three regions is connected.
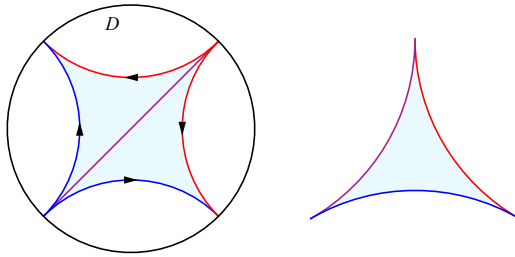
**Figure 1.** The thrice-punctured sphere can be obtained from the Poincaré disc ($D$) model of the hyperbolic plan by identifying the two ideal triangles as indicated: the purple side is already identified, blue is glued to blue, and red to red, according to the orientation given.

## 3. *n*-punctured spheres

The hyperbolic surfaces we will primarily focus on throughout this paper are *n*-punctured spheres, mainly because they are at once both simple (having cusps but no handles) and interesting. Proposition 3.5 gives the total area of an *n*-punctured sphere. Proposition 3.6 shows that for a certain range of areas, perimeter-minimizing double bubbles on an *n*-punctured sphere have disconnected boundary, a deviation from the topological properties of the standard double bubble.

**Definition 3.1.** An *n-punctured sphere* is constructed by doubling an ideal *n*-gon in hyperbolic 2-space and identifying the boundary.

The *n*-punctured sphere admits a hyperbolic metric for $n \geq 3$, so we assume henceforth that $n \geq 3$. Figure 1 gives an example of this construction in the case of the thrice-punctured sphere.

We have the following helpful proposition on single bubbles on the *n*-punctured sphere.

**Proposition 3.2** [Adams and Morgan 1999, Theorem 2.2]. *For single bubbles on a punctured surface, least-perimeter P is less than or equal to area A with equality precisely for horocycles about cusps. Moreover, if $A < \pi$, then a minimizer consists of horocycles about an arbitrary collection of cusps.*

**Remark 3.3.** Adams and Morgan [1999] further show that in the case of the thrice-punctured sphere, the hypothesis of this proposition can be extended to $A \leq \pi$.

In the proofs of our results we will make use of the following well-known facts in this area.

**Remark 3.4.** A horocycle about a cusp has constant curvature 1 and its length is equal to the area of the cusp neighborhood.

**Proposition 3.5.** *The total area of the n-punctured sphere is $2(n-2)\pi$.*

*Proof.* The area of an ideal triangle in hyperbolic 2-space is $\pi$. Since an ideal $n$-gon can be triangulated into $n - 2$ ideal triangles, the area of the ideal $n$-gon is $(n - 2)\pi$. The $n$-punctured sphere is composed of two ideal $n$-gons glued together and thus has area $2(n - 2)\pi$. □

**Proposition 3.6.** *Given* $0 < A_1 \leq A_2 < \pi - A_1$, *the least-perimeter way to enclose and separate areas* $A_1, A_2$ *on the n-punctured sphere is horocycles around cusps.*

*Proof.* Assume to the contrary the perimeter is less than or equal to $A_1 + A_2$ and the regions have common boundary. Then the shared boundary can be eliminated with the remaining boundary enclosing the single area $A_1 + A_2$. By our assumption the length of the remaining boundary is strictly less than $A_1 + A_2$. Since $A_1 + A_2 < \pi$, this is a contradiction of Proposition 3.2. □

## 4. The thrice-punctured sphere

The thrice-punctured sphere is equipped with unique hyperbolic structure with area $2\pi$ and constant Gaussian curvature $-1$. These features make the thrice-punctured sphere an ideal surface on which to explore the properties of double bubbles. Conjecture 4.1 says that horocycles are perimeter-minimizing for small areas and that a $\theta$-curve is perimeter-minimizing for large areas, with the transition point for equal areas given by Proposition 4.2. Proposition 4.8 shows that for double bubbles with connected perimeter, all three regions must contain a cusp. Proposition 4.9 further restricts the topology.

**Conjecture 4.1.** *Given two areas* $0 < A_1 \leq A_2 \leq 2\pi - A_1 - A_2$, *a perimeter-minimizing double bubble on the thrice-punctured sphere consists of*

(1) *horocycles around cusps if* $A_1$ *is relatively small,*

(2) *a $\theta$-curve with each region containing one cusp (*unique up to the three-fold symmetry*) if $A_1$ is relatively large (*see Figure 2*).*
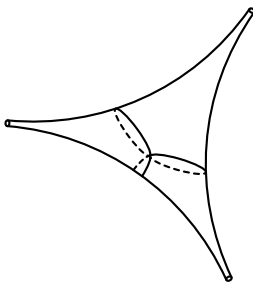


**Figure 2.** $\theta$-curves as pictured are conjectured to minimize perimeter for relatively large pairs of areas.

**Proposition 4.2.** *There exists a constant $A_0 \approx 1.7038$ such that given $0 < A_1 = A_2 \leq \frac{2\pi}{3}$, the symmetric $\theta$-curve enclosing areas $A_1$, $A_2$ is shorter than horocycles (of length $A_1 + A_2$) if and only if $A_1 = A_2 > A_0$.*

*Proof.* Let $H = \{x + yi \in \mathbb{C} \mid y > 0\}$ with the metric $ds = \sqrt{dx^2 + dy^2}/y$; this is the upper half-plane model of hyperbolic space. The length of a parametrized curve $\sigma : [a, b] \rightarrow H$ is given by

$$\text{length} = \int_a^b \frac{|\sigma'(t)|}{y(t)} \, dt.$$

The area of a region $R$ is given by

$$\text{area} = \iint_R \frac{1}{y} \, dx \, dy.$$

We consider the following construction in $H$. The thrice-punctured sphere can be considered as the quotient of two ideal triangles $H$ (the edges of these triangles are shown in blue in Figure 3 with the edges $e$ and $f$ being identified with $e'$ and $f'$ as shown). For computational ease we choose the radii of the semicircles $f$ and $f'$ to be 1.

Given $A_1 = A_2 = \frac{2\pi}{3}$, consider the pink $\theta$-curve $\phi$ of Figure 3, composed of three geodesics which each contain a cusp and meet at angles of $\frac{2\pi}{3}$. In the upper half-plane this curve consists of four circular arcs of radius 2 and angle $\frac{\pi}{6}$ and two vertical segments. Each of the arcs is centered at a vertex of the ideal triangle and runs from a vertical edge toward the center, while the two vertical segments run from the intersections of the arcs to the edges $f$ and $f'$.

By symmetry this curve divides the thrice-punctured sphere into three equal parts each having area $\frac{2\pi}{3}$. Due to symmetry the length of $\phi$ is $6l$, where $l$ is the length of just one of the vertical segments. Computing the length of the segment from $(1, \sqrt{3})$ to $(1, 1)$ using the formula given we obtain $l = \ln \sqrt{3} - \ln 1 = \frac{1}{2} \ln 3$. Thus the length of $\phi$ is $3 \ln 3$.

For $A_1 = A_2 = \frac{2\pi}{3}$, the $\theta$-curve has length $3 \ln 3 < \frac{4\pi}{3} = A_1 + A_2$, while for $A_1 = A_2 < \pi$, the horocycles of length $A_1 + A_2$ are minimizing by Proposition 3.6. Moreover, as $A_1 = A_2$ decreases, the symmetric $\theta$-curve gets longer and the horocycles get shorter. Therefore there is a constant $\pi < A_0 < \frac{2\pi}{3}$ such that the $\theta$-curve is shorter if and only if $A_1 = A_2 > A_0$.

Using Mathematica we were able to find an approximate value of $A_0$. For $A_1 = A_2 < \frac{2\pi}{3}$, we consider the same construction as for $\phi$, but shift it downwards a euclidean distance of $p$ to the red curve in Figure 3. This is the only possible $\theta$-curve enclosing $A_1$ and $A_2$ which satisfies the regularity and constant curvature conditions of a perimeter-minimizing double bubble. By symmetry, the length is given by adding four times the length of one arc (we take the one centered at $(0, p)$) to two
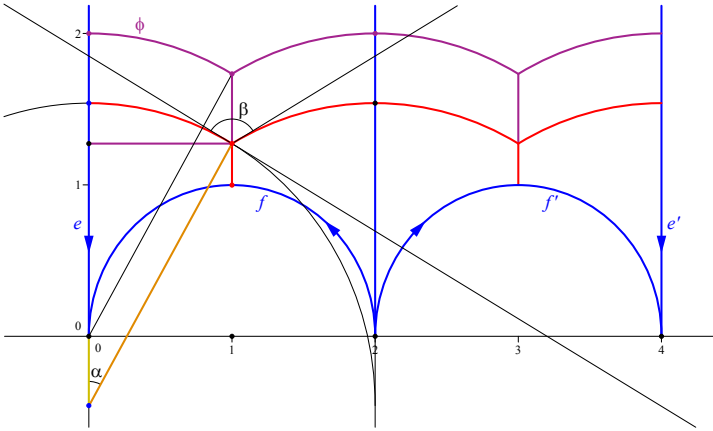
**Figure 3.** A $\theta$-curve on a thrice-punctured sphere in the upper half-plane allows us to parametrize the area and perimeter of the $\theta$-curve for equal areas.

times the length of one vertical segment (we take the one starting at $(1, 1)$). Using the standard parametrizations for these curves and the length formula given we obtain

$$\text{Perimeter}(p) = 4 \int_{\frac{\pi}{3}}^{\frac{\pi}{2}} \frac{1}{\sin x - \frac{p}{2}} \, dx - 2 \ln(\sqrt{3} - p).$$

The area enclosed by the red curve is given by taking four times area of the region between the red arc and the first half of $f$. Applying the given formula for computing area we have

$$\text{Area}(p) = 4 \int_0^1 \int_{\sqrt{1-(x-1)^2}}^{\sqrt{4-x^2}-p} \frac{1}{y} \, dy \, dx = 4 \int_0^1 -\frac{1}{\sqrt{4-x^2}-p} + \frac{1}{1-(x-1)^2} \, dx.$$

Given these parametrizations of area and perimeter, we can plot the perimeter against the area as in Figure 4. Further computations via Mathematica show that a $\theta$-curve is more efficient than horocycles for areas greater than about $3.4076/2$. $\square$

**Lemma 4.3.** *In the hyperbolic plane, for a disc of area $A$ and perimeter $P$ the following statements hold*:

(1) *If $A \leq \pi$, then $P \geq 2.2A$.*

(2) *If $A \leq \frac{\pi}{2}$, then $P \geq 3A$.*

(3) *If $A \leq \frac{4\pi}{9}$, then $P \geq \sqrt{10}$.*

(4) *If $A \leq \frac{4\pi}{15}$, then $P \geq 4A$.*

(5) *If $A < 8\pi/(9 + 3\sqrt{13})$, then $P > \frac{4\pi}{3}$.*
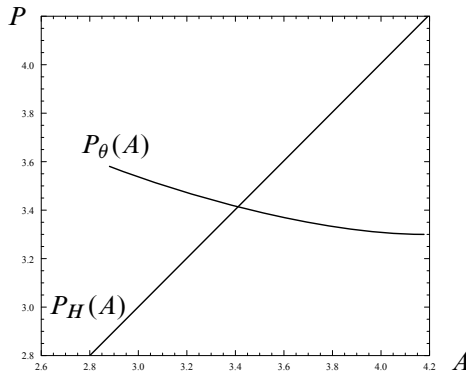
**Figure 4.** The $\theta$-curve is shorter than horocycles for equal areas greater than about 1.7.

*Proof.* Set $c = P/A$. If we parametrize area and perimeter of such a disc using the hyperbolic radius, $s$, then

$$c = \frac{2\pi \sinh s}{4\pi \sinh^2 \frac{s}{2}} = \coth \frac{s}{2}.$$

Notice that $\coth \frac{s}{2}$ is decreasing with $s$, whereas $A = 4\pi \sinh^2 \frac{s}{2}$ is increasing with $s$. Therefore $\coth \frac{s}{2}$ is bounded below by its value at the hyperbolic radius corresponding to the largest $A$. Suppose $A \leq \pi$. We solve $A \leq 4\pi \sinh^2 \frac{s}{2}$ to find $s \leq \cosh^{-1} \frac{3}{2}$. Hence $c \leq \coth\left(\frac{1}{2} \cosh^{-1}\left(\frac{3}{2}\right)\right) \approx 2.22$. Thus $P = cA \geq 2.2A$. Therefore, the first statement holds. Statements (2)–(4) are shown by the same method.

To show (5), we suppose that $A > 8\pi/(9 + 3\sqrt{13})$. Since $P(A) = \sqrt{A^2 + 4\pi A}$ is strictly increasing for all positive $A$, we have that for $A > 8\pi/(9 + 3\sqrt{13})$,

$$P > \sqrt{\left(\frac{8\pi}{9 + 3\sqrt{13}}\right)^2 + 4\pi \frac{8\pi}{9 + 3\sqrt{13}}} = \frac{4\pi}{3}. \qquad \square$$

**Remark 4.4.** In Lemma 4.3(1)–(4) both inequalities of each statement may be made strict and the statements will still hold. The method of proof is the same.

**Lemma 4.5.** *For two regions on the thrice-punctured sphere with areas $A_1$ and $A_2$ such that $A_1, A_2 \leq A_3 = 2\pi - A_1 - A_2$, we have that $A_1, A_2 \leq \pi$.*

*Proof.* If this was not true, the total area $A_1 + A_2 + A_3$ would exceed $2\pi$, which is the area of the thrice-punctured sphere (Proposition 3.5). $\qquad \square$

**Lemma 4.6.** *Given a double bubble with regions of areas $0 < A_1, A_2 \leq 2\pi - A_1 - A_2$ and perimeters $P_i$, the total perimeter $P$ satisfies $P \geq A_1 + \frac{1}{2} P_2$.*

*Proof.* Denote the area and perimeter of the complementary region by $A_3$ and $P_3$. By Lemma 4.5, $A_1 \leq \pi$. Thus Proposition 3.2 implies that $P_1 \geq A_1$. If $A_3 < \pi$, then

$$P_3 \geq A_3 = 2\pi - A_1 - A_2 \geq (2A_1 + A_2) - A_1 - A_2 = A_1.$$

If $A_3 > \pi$, then

$$P_3 \geq 2\pi - A_3 = 2\pi - (2\pi - A_1 - A_2) = A_1 + A_2 \geq A_1.$$

Therefore the total perimeter satisfies

$$P = \tfrac{1}{2}(P_1 + P_2 + P_3) \geq \tfrac{1}{2}(2A_1 + P_2) = A_1 + \tfrac{1}{2}P_2. \qquad \square$$

**Proposition 4.7.** *On a thrice-punctured sphere, a curve enclosing and separating regions $R_i$ of perimeters $P_i$ and areas $A_1$, $A_2 \leq 2\pi - A_1 - A_2$ has total perimeter $P > A_1 + A_2$ if $R_1$ or $R_2$ is a union of topological discs. In particular, it is not perimeter-minimizing.*

*Proof.* Suppose $R_2$ is the union of topological discs. Let $P_i$ denote the perimeter of $R_i$. Since the disc is isoperimetric in the hyperbolic plane, $P_2$ is greater than or equal to the perimeter of a hyperbolic disc of the same area. By Lemma 4.5, $A_2 \leq \pi$. Thus, by Lemma 4.3(1), $P_2 \geq 2.2A_2$. By Lemma 4.6, the total perimeter $P$ satisfies

$$P \geq A_1 + \tfrac{1}{2}P_2 > A_1 + \tfrac{1}{2}(2.2)A_2 > A_1 + A_2.$$

Therefore it cannot be perimeter-minimizing, because horocycles on two separate cusps have perimeter $A_1 + A_2$. $\qquad \square$

**Proposition 4.8.** *In a perimeter-minimizing double bubble with connected perimeter containing regions $R_i$ of perimeters $P_i$ and areas $A_1$, $A_2 \leq A_3 = 2\pi - A_1 - A_2$, all three regions contain a cusp.*

*Proof.* Both regions must have a component which is not a topological disc; otherwise horocycles enclosing the same area would be shorter than the perimeter of our double bubble by Proposition 4.7, contradicting the fact that our bubble is perimeter-minimizing. These components of regions which aren't topological discs must contain cusps (they can't be annular regions since the perimeter is connected).

Suppose that $R_3$ is the union of topological discs. Then $P_3$ is greater than or equal to the perimeter of the hyperbolic disc of area $A_3$. Since $dP/dA$ of the hyperbolic disc is always positive and $A_3 \geq \frac{2\pi}{3}$, $P_3$ is greater than or equal to the perimeter of the hyperbolic disc of area $\frac{2\pi}{3}$, which is $\frac{2\sqrt{7}\pi}{3}$. Therefore we have $P > P_3 > \frac{2\sqrt{7}\pi}{3} > \frac{4\pi}{3} \geq A_1 + A_2$, a contradiction. $\qquad \square$

**Proposition 4.9.** *Consider a double bubble enclosing areas $0 < A_1, A_2 \leq 2\pi - A_1 - A_2$, consisting of four region components $C_i$ of areas $A_1 - a_1$, $a_1$, $a_2$, $A_2 - a_2$, where $a_1 \leq a_2$, and each $C_i$ is adjacent only to $C_{i-1}$ and $C_{i+1}$ for $1 < i < 4$. Suppose*
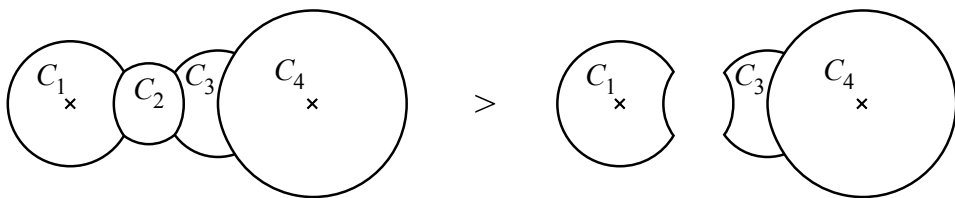
**Figure 5.** Lower bound on perimeter of four components of regions with two cusps.

*that $C_1$ and $C_4$ have no common boundary and are each connected and contain cusps, and that the union of $C_2$ and $C_3$ (not necessarily connected) is the union of topological discs. Then the total perimeter satisfies $P > A_1 + A_2$, and the double bubble is not perimeter-minimizing.*

*Proof.* Suppose that $\left(\frac{\sqrt{10}}{2} - 1\right)a_2 \le a_1$ and $\frac{4\pi}{15} \le a_2$. Then

$$a_1 + a_2 \ge \frac{\sqrt{10}}{2}a_2 \ge \frac{\sqrt{10}}{2}\frac{4\pi}{15} > \frac{8\pi}{3(3 + \sqrt{13})}.$$

Therefore $a_1 + a_2 > 8\pi/(3(3 + \sqrt{13}))$. We conclude that at least one of the following conditions must be satisfied:

(1) $8\pi/(3(3 + \sqrt{13})) < a_1 + a_2$.

(2) $a_1 \le \left(\frac{\sqrt{10}}{2} - 1\right)a_2$ and $a_2 < \frac{4\pi}{9}$.

(3) $a_2 < \frac{4\pi}{15}$.

Therefore it suffices to show $P > A_1 + A_2$ for the three cases where at least one of these conditions is satisfied.

<u>Case 1</u>: Since the union of $C_2$ and $C_3$ is the union of topological discs with boundary and the disc is isoperimetric in the hyperbolic plane, the length of the boundary of their union is greater than the perimeter of a hyperbolic disc of area $a_1 + a_2$. Therefore, by Lemma 4.3(5), $P > \frac{4\pi}{3} > A_1 + A_2$.

To show the remaining cases, we remove the unshared perimeter of $C_2$ (see Figure 5) and consider the sum of $P_1$ and the total perimeter of $C_3$ and $C_4$. Since $A_1 - a_1 \le A_1 \le \pi$ (Lemma 4.5), by Proposition 3.2, $P_1 \ge A_1 - a_1$. Since $C_3$ is the union of topological discs, the total perimeter of $C_3$ and $C_4$ is bounded below by $A_2 - a_2 + \frac{1}{2}P_3$, by Lemma 4.6. Thus $P \ge A_1 - a_1 + A_2 - a_2 + \frac{1}{2}P_3$.

<u>Case 2</u>: Since $a_2 < \frac{4\pi}{9}$, by Lemma 4.3(2) we have $P_3 > \sqrt{10}a_2$; thus

$$P \ge A_1 - a_1 + A_2 - a_2 + \frac{1}{2}P_3 > A_2 - a_2 + A_1 - a_1 + \frac{\sqrt{10}}{2}a_2.$$

Since $a_1 \leq \left(\frac{\sqrt{10}}{2} - 1\right)a_2$, we have

$$P > A_2 - a_2 + A_1 - a_1 + \left(\tfrac{\sqrt{10}}{2} - 1\right)a_2 + a_2$$
$$\geq A_2 - a_2 + A_1 - a_1 + a_1 + a_2 = A_1 + A_2.$$

<u>Case 3</u>:  By Lemma 4.3(3), we have $P_3 > 4a_2$; thus

$$P \geq A_2 - a_2 + A_1 - a_1 + \tfrac{1}{2}P_3 > A_2 - a_2 + A_1 - a_1 + 2a_2$$
$$\geq A_2 - a_2 + A_1 - a_1 + a_2 + a_1 = A_1 + A_2.$$

We conclude that $P > A_1 + A_2$. Hence it is not perimeter-minimizing, as horocycles on separate cusps have perimeter $A_1 + A_2$.    □

## 5. Once-punctured surfaces

Some of the methods employed in Section 4, can be applied to other hyperbolic surfaces of constant Gaussian curvature $-1$ that share some features of the thrice-punctured sphere, such as having area of $2\pi$ and at least one cusp, but lack its fixed hyperbolic structure. For example, a once punctured torus has many hyperbolic structures, yet all have area $2\pi$. Proposition 5.1 shows that for relatively small areas on such a surface, two horocycles have less perimeter than one horocycle with a lens.

**Proposition 5.1.** *Given two areas* $0 < A_1, A_2 \leq \frac{4\pi}{15}$ *on a punctured surface of area* $2\pi$, *the union of two horocycles about the cusp enclosing and separating* $A_1$ *and* $A_2$ *is shorter than a horocycle with a lens.*

*Proof.* Without loss of generality suppose that $A_1$ is not on the cusp. Since $A_1 \leq \pi$ (Lemma 4.5), by Proposition 3.2, our surface has the same isoperimetric profile for single bubbles as the thrice-punctured sphere. Thus Lemma 4.6 holds, and the total perimeter, $P$, of our enclosure satisfies the inequality $P \geq A_2 + \frac{1}{2}P_1$. By Lemma 4.3(4), $P_1 \geq 4A_1$ for $A_1 \geq \frac{4\pi}{15}$. Thus $P \geq A_2 + 2A_1$.    □

## Acknowledgements

## References

[Adams and Morgan 1999]  C. Adams and F. Morgan, "Isoperimetric curves on hyperbolic surfaces", *Proc. Amer. Math. Soc.* **127**:5 (1999), 1347–1356. MR Zbl

[Cotton and Freeman 2002] A. Cotton and D. Freeman, "The double bubble problem in spherical space and hyperbolic space", *Int. J. Math. Math. Sci.* **32**:11 (2002), 641–699. MR Zbl

[Foisy 1991] J. Foisy, "Soap bubble clusters in $\mathbb{R}^2$ and $\mathbb{R}^3$", undergraduate thesis, 1991.

[Foisy et al. 1993] J. Foisy, M. Alfaro, J. Brock, N. Hodges, and J. Zimba, "The standard double soap bubble in $\mathbb{R}^2$ uniquely minimizes perimeter", *Pacific J. Math.* **159**:1 (1993), 47–59. MR Zbl

[Hass et al. 1995] J. Hass, M. Hutchings, and R. Schlafly, "The double bubble conjecture", *Electron. Res. Announc. Amer. Math. Soc.* **1**:3 (1995), 98–102. MR Zbl

[Howards et al. 1999] H. Howards, M. Hutchings, and F. Morgan, "The isoperimetric problem on surfaces", *Amer. Math. Monthly* **106**:5 (1999), 430–439. MR Zbl

[Hutchings et al. 2002] M. Hutchings, F. Morgan, M. Ritoré, and A. Ros, "Proof of the double bubble conjecture", *Ann. of Math.* (2) **155**:2 (2002), 459–489. MR Zbl

[Morgan 1994] F. Morgan, "Soap bubbles in $\mathbb{R}^2$ and in surfaces", *Pacific J. Math.* **165**:2 (1994), 347–361. MR Zbl

wbb1@williams.edu                Department of Mathematics and Statistics,
                                 Williams College, Williamstown, MA, United States

bcb02011@mymail.pomona.edu       Department of Mathematics, Pomona College,
                                 Claremont, CA, United States

aloving@hawaii.edu               Alyssa Loving, Department of Mathematics,
                                 University of Hawaii at Hilo, Hilo, HI, United States

setammen@uga.edu                 Department of Mathematics, University of Georgia,
                                 Athens, GA, United States

# What is odd about binary Parseval frames?

Zachery J. Baker, Bernhard G. Bodmann, Micah G. Bullock,
Samantha N. Branum and Jacob E. McLaney

(Communicated by David Royal Larson)

This paper examines the construction and properties of binary Parseval frames. We address two questions: When does a binary Parseval frame have a complementary Parseval frame? Which binary symmetric idempotent matrices are Gram matrices of binary Parseval frames? In contrast to the case of real or complex Parseval frames, the answer to these questions is not always affirmative. The key to our understanding comes from an algorithm that constructs binary orthonormal sequences that span a given subspace, whenever possible. Special regard is given to binary frames whose Gram matrices are circulants.

## 1. Introduction

Much of the literature on frames, from its beginnings in nonharmonic Fourier analysis [Duffin and Schaeffer 1952] to comprehensive overviews of theory and applications [Christensen 2003; Kovačević and Chebira 2007a; 2007b], assumes an underlying structure of a real or complex Hilbert space to study approximate expansions of vectors. Indeed, the correspondence between vectors in Hilbert spaces and linear functionals given by the Riesz representation theorem provides a convenient way to characterize Parseval frames, sequences of vectors that behave in a way that is similar to orthonormal bases without requiring the vectors to be linearly independent [Christensen 2003]. Incorporating linear dependence relations is useful to permit more flexibility for expansions and to suppress errors that may model faulty signal transmissions in applications [Marshall 1984; 1989; Rath and Guillemot 2003; 2004; Holmes and Paulsen 2004; Puschel and Kovačević 2005; Bodmann and Paulsen 2005].

The concept of frames has also been established even in vector spaces without a positive definite inner product [Bodmann et al. 2009; Han et al. 2007]. In fact, the

well-known theory of binary codes can be seen as a form of frame theory in which linear dependence relations among binary vectors are examined [MacWilliams and Sloane 1977; Haemers et al. 1999; Betten et al. 2006]. Here, binary vector spaces are defined over the finite field with two elements; a frame for a finite-dimensional binary vector space is simply a spanning sequence [Bodmann et al. 2009]. In a preceding paper [Bodmann et al. 2014], the study of binary codes from a frame-theoretic perspective has led to additional combinatorial insights in the design of error-correcting codes.

The present paper is concerned with binary Parseval frames. These binary frames provide explicit expansions of binary vectors using a bilinear form that resembles the dot product in Euclidean spaces. In contrast to the inner product on real or complex Hilbert spaces, there are many nonzero vectors whose dot product with themselves vanishes. Such vectors have special significance in our results. Due to the number of nonzero entries they contain, we call them *even* vectors, and if a vector is not even, we call it *odd*. As a consequence of the degeneracy of the bilinear form, there are some striking differences with frame theory over real or complex Hilbert spaces. In this paper, we explore the construction and properties of binary Parseval frames, and compare them with real and complex ones. Our main results are as follows.

In the real or complex case, it is known that each Parseval frame has a Naimark complement [Christensen 2003; Han and Larson 2000]. The complementarity is most easily formulated by stating that the Gram matrices of two complementary Parseval frames sum to the identity. We show that in the binary case, not every Parseval frame has a Naimark complement. We also show that a necessary and sufficient condition for its existence is that the Parseval frame contains at least one even vector.

Moreover, we study the structure of Gram matrices. The Gram matrices of real or complex Parseval frames are characterized as symmetric or hermitian idempotent matrices. The binary case requires the additional condition that at least one column vector of the matrix is odd.

The general results we obtain are illustrated with examples. Special regard is given to cyclic binary Parseval frames, whose Gram matrices are circulants.

## 2. Preliminaries

We define the notions of a binary frame and a binary Parseval frame as in a previous paper [Bodmann et al. 2009]. The vector space that these sequences of vectors span is the direct sum $\mathbb{Z}_2^n = \mathbb{Z}_2 \oplus \cdots \oplus \mathbb{Z}_2$ of $n$ copies of $\mathbb{Z}_2$ for some $n \in \mathbb{N}$. Here, $\mathbb{Z}_2$ is the field of binary numbers with the two elements 0 and 1, the neutral element with respect to addition and the multiplicative identity, respectively.

**Definition 2.1.** A *binary frame* is a sequence $\mathcal{F} = \{f_1, \ldots, f_k\}$ in a binary vector space $\mathbb{Z}_2^n$ such that span $\mathcal{F} = \mathbb{Z}_2^n$.

A simple example of a frame is the canonical basis $\{e_1, e_2, \ldots, e_n\}$ for $\mathbb{Z}_2^n$. The $i$-th vector has components $(e_i)_j = \delta_{i,j}$, $j \in \{1, 2, \ldots, n\}$, and thus $(e_i)_i = 1$ is the only nonzero entry for $e_i$. Consequently, a vector $x = (x_i)_{i=1}^n$ is expanded in terms of the canonical basis as $x = \sum_{i=1}^n x_i e_i$.

Frames provide similar expansions of vectors in linear combinations of the frame vectors. Parseval frames are especially convenient for this purpose because the linear combination can be determined with little effort. In the real or complex case, this only requires computing values of inner products between the vector to be expanded and the frame vectors. Although we cannot introduce a nondegenerate inner product in the binary case, we define Parseval frames using a bilinear form that resembles the dot product on $\mathbb{R}^n$. Other choices of bilinear forms and a more general theory of binary frames have been investigated elsewhere; see [Hotovy et al. 2015].

**Definition 2.2.** The *dot product* on $\mathbb{Z}_2^n$ is the bilinear map $(\,\cdot\,,\cdot\,) : \mathbb{Z}_2^n \times \mathbb{Z}_2^n \to \mathbb{Z}_2$ given by

$$\left( \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \right) := \sum_{i=1}^n x_i y_i.$$

With the help of this dot product, we define a Parseval frame for $\mathbb{Z}_2^n$.

**Definition 2.3.** A *binary Parseval frame* is a sequence of vectors $\mathcal{F} = \{f_1, \ldots, f_k\}$ in $\mathbb{Z}_2^n$ such that, for all $x \in \mathbb{Z}_2^n$, the sequence satisfies the reconstruction identity

$$x = \sum_{j=1}^k (x, f_j) f_j. \tag{2-1}$$

To keep track of the specifics of such a Parseval frame, we then also say that $\mathcal{F}$ is a *binary $(k, n)$-frame*.

In the following, we use matrix algebra whenever it is convenient for establishing properties of frames. We write $A \in M_{m,n}(\mathbb{Z}_2)$ when $A$ is an $m \times n$ matrix with entries in $\mathbb{Z}_2$ and identify $A$ with the linear map from $\mathbb{Z}_2^n$ to $\mathbb{Z}_2^m$ induced by left multiplication of any (column) vector $x \in \mathbb{Z}_2^n$ with $A$. We let $A^*$ denote the adjoint of $A \in M_{m,n}(\mathbb{Z}_2)$; that is, $(Ax, y) = (x, A^*y)$ for all $x \in \mathbb{Z}_2^n$, $y \in \mathbb{Z}_2^m$ and consequently, $A^*$ is the transpose of $A$.

**Definition 2.4.** Each frame $\mathcal{F} = \{f_1, \ldots, f_k\}$ is associated with its *analysis matrix* $\Theta_{\mathcal{F}}$, whose $i$-th row is given by the $i$-th frame vector for $i \in \{1, 2, \ldots, k\}$. Its transpose $\Theta_{\mathcal{F}}^*$ is called the *synthesis matrix*.

With the help of matrix multiplication, the reconstruction formula (2-1) of a binary $(k, n)$-frame $\mathcal{F}$ with analysis matrix $\Theta_{\mathcal{F}}$ is simply expressed as

$$\Theta_{\mathcal{F}}^* \Theta_{\mathcal{F}} = I_n, \tag{2-2}$$

where $I_n$ is the $n \times n$ identity matrix. We also note that for any $x, y \in \mathbb{Z}_2^n$, their dot product is unchanged by applying $\Theta_{\mathcal{F}}$,

$$(\Theta_{\mathcal{F}} x, \Theta_{\mathcal{F}} y) = (x, y),$$

which motivates speaking of $\Theta_{\mathcal{F}}$ as an *isometry*, as in the case of real or complex inner product spaces.

Another way to interpret identity (2-2) is in terms of the *column* vectors of $\Theta_{\mathcal{F}}$. Again borrowing a concept from Euclidean spaces, we introduce orthonormality.

**Definition 2.5.** We say that a sequence of vectors $\{v_1, v_2, \ldots, v_r\}$ in $\mathbb{Z}_2^n$ is *orthonormal* if $(v_i, v_j) = \delta_{i,j}$ for $i, j \in \{1, 2, \ldots, r\}$; that is, the dot product of the pair $v_i$ and $v_j$ vanishes unless $i = j$, in which case it is equal to 1.

Inspecting the matrix identity (2-2), we see that a binary $k \times n$ matrix $\Theta$ is the analysis matrix of a binary Parseval frame if and only if the columns of $\Theta$ form an orthonormal sequence in $\mathbb{Z}_2^k$.

The orthogonality relations between the frame vectors are recorded in the Gram matrix, whose entries consist of the dot products of all pairs of vectors.

**Definition 2.6.** The *Gram matrix* of a binary frame $\mathcal{F} = \{f_1, f_2, \ldots, f_k\}$ for $\mathbb{Z}_2^n$ is the $k \times k$ matrix $G$ with entries $G_{i,j} = (f_j, f_i)$.

It is straightforward to verify that the Gram matrix of $\mathcal{F}$ is expressed as the composition of the analysis and synthesis matrices,

$$G = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*.$$

The identity (2-2) implies that the Gram matrix of a Parseval frame satisfies

$$G = G^* = G^2.$$

For frames over the real or complex numbers, these equations characterize the set of all Gram matrices of Parseval frames as orthogonal projection matrices. However, in the binary case, this is only a necessary condition, as shown in the following proposition and the subsequent example.

**Proposition 2.7.** *If $M$ is binary matrix that satisfies $M = M^2 = M^*$ and it has only even column vectors, then $M$ is not the Gram matrix of a binary Parseval frame.*

*Proof.* If $G$ is the Gram matrix of a Parseval frame with analysis operator $\Theta$, then $G\Theta = \Theta\Theta^*\Theta = \Theta$, and thus for each column $\omega$ of $\Theta$, we obtain the eigenvector equation $G\omega = \omega$. By the orthonormality of the columns of $\Theta$, each $\omega$ is odd.

On the other hand, if $M$ has only even columns, then any eigenvector corresponding to eigenvalue 1 is even, because it is a linear combination of the column vectors of $M$. This means $M$ cannot be the Gram matrix of a binary Parseval frame. $\square$

The following example shows that idempotent symmetric matrices that are not Gram matrices of binary Parseval frames exist for any odd dimension $k \geq 3$.

**Example 2.8.** Let $k \geq 3$ be odd and let $M$ be the $k \times k$ matrix whose entries are all equal to 1 except for vanishing entries on the diagonal, $M_{i,j} = 1 - \delta_{i,j}$, $i, j \in \{1, 2, \ldots, k\}$. This matrix satisfies $M = M^2 = M^*$, but only has even columns, and by the preceding proposition, it is not the Gram matrix of a binary Parseval frame.

As shown in Section 4, having only even column vectors is the only way a binary symmetric idempotent matrix can fail to be the Gram matrix of a Parseval frame. The construction of Example 2.8 is intriguing because the alternative choice where $k$ is odd and all entries of $M$ are equal to 1 *is* the Gram matrix of a binary Parseval frame. The relation between these two alternatives can be interpreted as complementarity, which will be explored in more detail in the next section.

## 3. Complementarity for binary Parseval frames

Over the real or complex numbers, each Parseval frame has a so-called Naimark complement [Christensen 2003], also called a strong complement [Han and Larson 2000]; if $G$ is the Gram matrix of a real or complex Parseval frame, then it is an orthogonal projection matrix, and so is $I - G$, which makes it the Gram matrix of a complementary Parseval frame.

We adopt the same definition for the binary case.

**Definition 3.1.** Two binary Parseval frames $\mathcal{F}$ and $\mathcal{G}$ having analysis operators $\Theta_{\mathcal{F}} \in M_{k,n}(\mathbb{Z}_2)$ and $\Theta_{\mathcal{G}} \in M_{k,k-n}(\mathbb{Z}_2)$ are *complementary* if

$$\Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^* + \Theta_{\mathcal{G}} \Theta_{\mathcal{G}}^* = I_k.$$

We also say that $\mathcal{F}$ and $\mathcal{G}$ are *Naimark complements* of each other.

There is an equivalent statement of complementarity in terms of the block matrix $U = (\Theta_{\mathcal{F}} \, \Theta_{\mathcal{G}})$, formed by adjoining $\Theta_{\mathcal{F}}$ and $\Theta_{\mathcal{G}}$, being *orthogonal*, meaning $UU^* = U^*U = I$, just as in the real case (or as $U$ being unitary in the complex case).

**Proposition 3.2.** *Two binary Parseval frames $\mathcal{F}$ and $\mathcal{G}$ having analysis operators $\Theta_{\mathcal{F}} \in M_{k,n}(\mathbb{Z}_2)$ and $\Theta_{\mathcal{G}} \in M_{k,k-n}(\mathbb{Z}_2)$ are complementary if and only if the block matrix $(\Theta_{\mathcal{F}} \, \Theta_{\mathcal{G}})$ is an orthogonal $k \times k$ matrix.*

*Proof.* In terms of the block matrix $(\Theta_{\mathcal{F}} \, \Theta_{\mathcal{G}})$, the complementarity is expressed as

$$(\Theta_{\mathcal{F}} \, \Theta_{\mathcal{G}})(\Theta_{\mathcal{F}} \, \Theta_{\mathcal{G}})^* = I_k \, .$$

Since $U = (\Theta_{\mathcal{F}} \, \Theta_{\mathcal{G}})$ is a square matrix, $UU^* = I$ is equivalent to $U^*$ also being a left inverse of $U$, meaning $UU^* = U^*U = I_k$, and thus $U$ is orthogonal. $\square$

In the binary case, not every Parseval frame has a Naimark complement. For example, if $k \geq 3$ is odd and $n = 1$, the frame consisting of $k$ vectors $\{1, 1, \ldots, 1\}$ in $\mathbb{Z}_2$ is Parseval, and the Gram matrix $G$ is the $k \times k$ matrix whose entries are all equal to 1. However, $I - G \equiv I + G$ is the matrix $M$ appearing in Example 2.8, which is not the Gram matrix of a binary Parseval frame. This motivates the search for a condition that characterizes the existence of complementary Parseval frames.

*A simple condition for the existence of complementary Parseval frames.* We observe that if $\mathcal{F}$ is a Parseval frame with analysis operator $\Theta_{\mathcal{F}}$ that extends to an orthogonal matrix, then the column vectors of $\Theta_{\mathcal{F}}$ are a subset of a set of $n$ orthonormal vectors. This is true in the binary as well as the real or complex case. Thus, one could try to relate the construction of a complementary Parseval frame to a Gram–Schmidt orthogonalization strategy. Indeed, this idea allows us to formulate a concrete condition that characterizes when $\mathcal{F}$ has a complementary Parseval frame. We prepare this result with a lemma about extending orthonormal sequences.

**Lemma 3.3.** *A binary orthonormal sequence $\mathcal{Y} = \{v_1, v_2, \ldots, v_r\}$ in $\mathbb{Z}_2^k$ with $r \leq k-1$ extends to an orthonormal sequence $\{v_1, v_2, \ldots, v_k\}$ if and only if $\sum_{i=1}^{r} v_i \neq \iota_k$, where $\iota_k$ is the vector in $\mathbb{Z}_2^k$ whose entries are all equal to 1.*

*Proof.* If the sequence extends, then $\{v_1, v_2, \ldots, v_k\}$ forms a Parseval frame for $\mathbb{Z}_2^k$, and by the orthonormality, $\sum_{i=1}^{k} v_i = \sum_{i=1}^{k} (\iota_k, v_i) v_i = \iota_k$. On the other hand, the orthonormality forces the set $\{v_1, v_2, \ldots, v_k\}$ to be linearly independent, so $\iota_k$ cannot be expressed as a linear combination of a proper subset.

To show the converse, we use an inductive proof. Let $V$ be the analysis operator associated with an orthonormal sequence $\{v_1, v_2, \ldots, v_s\}$, $r \leq s \leq k - 1$, satisfying $\sum_{i=1}^{s} v_i \neq \iota_k$. To extend the sequence by one vector, we need to find $v_{s+1}$ with $(v_{s+1}, v_{s+1}) = (v_{s+1}, \iota_k) = 1$ and with $(v_j, v_{s+1}) = 0$ for all $1 \leq j \leq s$. Using block matrices this is summarized in the equation

$$\begin{pmatrix} V \\ \iota_k^* \end{pmatrix} v_{s+1} = \begin{pmatrix} 0_s \\ 1 \end{pmatrix}, \tag{3-1}$$

where $0_s$ is the zero vector in $\mathbb{Z}_2^s$.

In order to verify that this equation is consistent, we note that by the orthonormality of the sequence $\{v_1, v_2, \ldots, v_s\}$, the vector $\iota_k$ is a linear combination if and only if $\sum_{i=1}^{s} v_i = \iota_k$. Thus, there exists $v_{s+1}$ which extends the orthonormal sequence. This is all that is needed if $s = k - 1$.

Next, we need to show that if $s \leq k - 2$, then a solution $v_{s+1}$ can be chosen so that $\sum_{i=1}^{s+1} v_i \neq \iota_k$, so that the iterative extension procedure can be continued. The solution set of (3-1) forms an affine subspace of $\mathbb{Z}_2^k$ having dimension $k - (s + 1)$, and thus contains $2^{k-s-1}$ elements. If $s \leq k-2$, there are at least two elements in this affine subspace. Consequently, there is one choice of $v_{s+1}$ such that $\sum_{i=1}^{s+1} v_i \neq \iota_k$. $\square$

We are ready to characterize the complementarity property for binary Parseval frames. The condition that determines the existence of a Naimark complement is whether at least one frame vector is even, that is, its entries sum to zero.

**Theorem 3.4.** *A binary $(k, n)$-frame $\mathcal{F}$ with $n < k$ has a complementary Parseval frame if and only if at least one frame vector is even.*

*Proof.* The existence of a complementary Parseval frame is by Proposition 3.2 equivalent to the sequence of column vectors $\{\omega_1, \omega_2, \ldots, \omega_n\}$ of $\Theta_{\mathcal{F}}$ having an extension to an orthonormal sequence of $k$ elements.

The condition that at least one frame vector is even can be stated as $\Theta_{\mathcal{F}} \iota_n \neq \iota_k$ or, expressed in terms of the column vectors, as $\sum_{i=1}^{n} \omega_n \neq \iota_k$.

The preceding lemma thus provides the existence of a complementary Parseval frame via the extension of $\{\omega_1, \omega_2, \ldots, \omega_n\}$ if and only if $\sum_{i=1}^{n} \omega_n \neq \iota_k$.          $\square$

***A catalog of binary Parseval frames with the complementarity property.*** A previous work contained a catalog of binary Parseval frames for $\mathbb{Z}_2^n$ when $n$ was small [Bodmann et al. 2009]. Here, we wish to compile a list of the binary Parseval frames that have a complementary Parseval frame. For notational convenience, we consider $\Theta_{\mathcal{F}}$ instead of the sequence of frame vectors. By Proposition 3.2, every such $\Theta_{\mathcal{F}}$ is obtained by a selection of columns from a binary orthogonal matrix, so we could simply list the set of all orthogonal matrices for small $k$. However, such a list quickly becomes extensive as $k$ increases. To reduce the number of orthogonal matrices, we note that although the frame depends on the order in which the columns are selected to form $\Theta_{\mathcal{F}}$, the Gram matrix does not. Identifying frames whose Gram matrices coincide has already been used to avoid repeating information when examining real or complex frames [Balan 1999] and binary frames [Bodmann et al. 2009]. We consider an even coarser underlying equivalence relation [Goyal et al. 2001; Holmes and Paulsen 2004; Bodmann and Paulsen 2005] that has also appeared in the context of binary frames [Bodmann et al. 2009].

**Definition 3.5.** Two families $\mathcal{F} = \{f_1, f_2, \ldots, f_k\}$ and $\mathcal{G} = \{g_1, g_2, \ldots, g_k\}$ in $\mathbb{Z}_2^n$ are called *switching equivalent* if there is an orthogonal $n \times n$ matrix $U$ and a permutation $\pi$ of the set $\{1, 2, \ldots, k\}$ such that

$$f_j = U g_{\pi(j)} \quad \text{for all } j \in \{1, 2, \ldots\}.$$

Representing the permutation $\pi$ by the associated permutation matrix $P$ with entries $P_{i,j} = \delta_{i,\pi(j)}$ gives that if $\mathcal{F}$ and $\mathcal{G}$ are switching equivalent, then $\Theta_{\mathcal{F}} = P \Theta_{\mathcal{G}} U$, where $U$ is an orthogonal $n \times n$ matrix and $P$ is a $k \times k$ permutation matrix. Alternatively, switching equivalence is stated in the form of an identity for the corresponding Gram matrices.

**Theorem 3.6** [Bodmann et al. 2009]. *Two binary $(k, n)$-frames $\mathcal{F}$ and $\mathcal{G}$ are switching equivalent if and only if their Gram matrices are related by conjugation with a $k \times k$ permutation matrix $P$,*

$$G_{\mathcal{F}} = P G_{\mathcal{G}} P^*.$$

We deduce a consequence for switching equivalence and Naimark complements, which is inferred from the role of the Gram matrices in the definition of complementarity.

**Corollary 3.7.** *If $\mathcal{F}$ and $\mathcal{G}$ are switching-equivalent binary $(k, n)$-frames, then $\mathcal{F}$ has a Naimark complement if and only if $\mathcal{G}$ does.*

Thus, to provide an exhaustive list, we only need to ensure that at least one representative of each switching equivalence class appears as a selection of columns in the orthogonal matrices we include. To reduce the number of representatives, we identify matrices up to row and column permutations.

**Definition 3.8.** Two matrices $A, B \in M_{k,k}(\mathbb{Z}_2)$ are called *permutation equivalent* if there are two permutation matrices $P_1, P_2 \in M_{k,k}(\mathbb{Z}_2)$ such that $A = P_1 B P_2^*$.

**Proposition 3.9.** *If $U_1$ and $U_2$ are permutation-equivalent binary orthogonal matrices, then each $(k, n)$-frame $\mathcal{F}$ formed by a sequence of $n$ columns of $U_1$ is switching equivalent to a $(k, n)$-frame $\mathcal{G}$ formed with columns of $U_2$.*

*Proof.* Without loss of generality, we can assume that the analysis matrix $\Theta_{\mathcal{F}}$ is formed by the first $n$ columns of $U_1$. By the equivalence of $U_1$ and $U_2$, we have $U_1 P_2 = P_1 U_2$ with permutation matrices $P_1$ and $P_2$. The right multiplication of $U_1$ with $P_2$ gives a column permutation, which identifies a sequence of columns in $P_1 U_2$ that is identical to the first $n$ columns of $U_1$. If $\mathcal{G}$ is obtained with the corresponding columns in $U_2$, then the Gram matrices of $\mathcal{F}$ and $\mathcal{G}$ are related by $G_{\mathcal{F}} = P_1 G_{\mathcal{G}} P_1^*$, which proves the switching equivalence.    □

A list of permutation-inequivalent orthogonal $k \times k$ matrices allows us to obtain the Gram matrix of each binary $(k, n)$-frame with a Naimark complement by selecting an appropriate choice of $n$ columns from an orthogonal $k \times k$ matrix to form $\Theta$ and then by applying a permutation matrix $P$ to obtain $G_{\mathcal{F}} = P \Theta \Theta^* P^*$.

Each representative of an equivalence class of orthogonal matrices can be chosen so that the columns are in lexicographical order. Table 1 contains a complete list of representatives of binary orthogonal matrices for $k \in \{3, 4, 5, 6\}$ from each permutation equivalence class. Each column vector in our list is recorded by the integer obtained from the binary expansion with the entries of the vector. For example, if a frame vector in $\mathbb{Z}_2^4$ is $f_1 = (1, 0, 1, 1)$, then it is represented by the integer $2^0 + 2^2 + 2^3 = 13$. Accordingly, in $\mathbb{Z}_2^4$, the standard basis is recorded as the sequence of numbers 1, 2, 4, 8.

| $k$ | nonequivalent $k \times k$ orthogonal matrices |
|---|---|
| 3 | (1, 2, 4) |
| 4 | (1, 2, 4, 8) <br> (7, 11, 13, 14) |
| 5 | (1, 2, 4, 8, 16) <br> (4, 11, 19, 25, 26) <br> (7, 8, 19, 21, 22) <br> (7, 11, 13, 14, 16) |
| 6 | (1, 2, 4, 8, 16, 32) <br> (4, 8, 19, 35, 49, 50) <br> (4, 11, 16, 35, 41, 42) <br> (4, 11, 19, 25, 26, 32) <br> (7, 8, 16, 35, 37, 38) <br> (7, 8, 19, 21, 22, 32) <br> (7, 11, 13, 14, 16, 32) <br> (13, 14, 28, 44, 55, 59) <br> (21, 22, 28, 47, 52, 59) <br> (25, 26, 28, 47, 55, 56) <br> (31, 37, 38, 44, 52, 59) <br> (31, 41, 42, 44, 55, 56) <br> (31, 47, 49, 50, 52, 56) <br> (31, 47, 55, 59, 61, 62) |

**Table 1.** All permutation-inequivalent binary orthogonal $k \times k$ matrices, $3 \leq k \leq 6$. Up to switching equivalence, the Gram matrix of each binary $(k, n)$-frame with a Naimark complement is obtained by selecting appropriate columns in one of the listed $k \times k$ orthogonal matrices.

## 4. Gram matrices of binary Parseval frames

The preceding section on complementarity hinged on the problem that even if $G$ is the Gram matrix of a binary Parseval frame, $I - G$ may not be, even though it is symmetric and idempotent. Again, there is a simple condition that needs to be added; Gram matrices of binary Parseval frames are symmetric and idempotent *and* have at least one odd column, that is, a column whose entries sum to 1. Because of the identity $G^2 = G$, having an odd column is equivalent to having a nonzero diagonal entry. Indeed, it has been shown that for *any* binary symmetric matrix $G$ without vanishing diagonal, there is a factor $\Theta$ such that $G = \Theta\Theta^*$ and the rank

of $\Theta$ is equal to that of $G$ [Lempel 1975]. The assumptions needed for our proof are stronger, but our algorithm for producing $\Theta$ appears to be more straightforward than the factorization procedure for general symmetric binary matrices.

**Theorem 4.1.** *A binary symmetric idempotent matrix $M$ is the Gram matrix of a Parseval frame if and only if it has at least one odd column.*

*Proof.* First, we re-express the condition on the columns of a symmetric $k \times k$ matrix $M$ in the equivalent form of the matrix $I_k + M$ having at least one even column or row. This, in turn, is equivalent to the inequality $(I_k + M)\iota_k \neq \iota_k$.

Next, we recall that both $M$ and $I_k + M$ are assumed to be idempotent. We observe that any vector $y \in \mathbb{Z}_2^k$ is in the range of an idempotent $P$ if and only if $Py = y$ if and only if $y$ is in the kernel of $I_k + P$.

Assuming $M$ is the Gram matrix of a Parseval frame, we have $M = \Theta\Theta^*$ where $\Theta$ has orthonormal columns and $(I_k + M)\Theta = 0$. Combining the two properties gives

$$\begin{pmatrix} I_k + M \\ \iota_k^* \end{pmatrix} \Theta = \begin{pmatrix} 0_{k,n} \\ \iota_n^* \end{pmatrix}.$$

This is inconsistent if and only if $\iota_k$ is in the span of the columns of the idempotent $I_k + M$, which is equivalent to $(I_k + M)\iota_k = \iota_k$.

Conversely, assuming that $M$ is symmetric and idempotent and has at least one odd column, we construct a matrix $\Theta$ with orthonormal columns such that $M = \Theta\Theta^*$.

We follow an inductive strategy similar to an earlier proof and construct an orthonormal sequence $\{\omega_1, \ldots, \omega_n\}$ as described in the following paragraph such that $n$ is the rank of $M$ and $M\omega_i = \omega_i$ for all $i \in \{1, 2, \ldots, n\}$. In that case, the range of $M$ is the span of the sequence, and so is the range of $M^*$. Moreover, if $\Theta$ contains the column vectors $\{\omega_1, \ldots, \omega_n\}$, then $M\omega_i = \omega_i = \Theta\Theta^*\omega_i$ implies $M = \Theta\Theta^*$ because the two matrices have rank at most $n$ and provide the identity map on the span of $n$ linearly independent vectors.

To begin with the induction, if $M$ has an odd column, then the fact that $M$ is idempotent gives that the equation $(I_k + M)\omega_1 = 0$ has this column vector as an odd solution. If this solution is unique, then $I_k + M$ has rank $k - 1$, $M$ has rank 1 and $\{\omega_1\}$ is the desired sequence. If the solution is not unique, then we can choose $\omega_1 \neq \iota_k$ and proceed with the induction.

Next, we extend a given orthonormal sequence $\{\omega_1, \ldots, \omega_s\}$ in the kernel of $I_k + M$ with $s \leq n - 2$ by one vector. Let $V$ be a matrix formed by a maximal set of linearly independent rows in $I_k + M$. Then if $M$ has rank $n$, the rank-nullity theorem gives that $V$ has $k - n$ rows. Letting $Y$ be the analysis matrix of the orthonormal sequence $\{\omega_1, \ldots, \omega_s\}$, extending it by one vector requires solving the equation

$$\begin{pmatrix} V \\ Y \\ \iota_k^* \end{pmatrix} \omega_{s+1} = \begin{pmatrix} 0_{k-n} \\ 0_s \\ 1 \end{pmatrix}. \tag{4-1}$$

In order to avoid producing an inconsistent equation during the induction process, we strengthen the induction assumption by the requirement that $\iota_k^*$ is not in the span of the rows of the matrix formed by $V$ and $Y$ and conclude in each step that $\iota_k^*$ is not in the span of the rows of the matrix formed by $V$ and $Y$ and $\omega_{s+1}^*$. As before, this is obtained by the fact that $VY^* = 0$, so if $\iota_k = \sum_{i=1}^{s+1} c_i \omega_i + v$ with $v$ being in the span of the columns of $V^*$, then $Yv = 0$ and orthonormality forces $c_i = 1$ for all $i \in \{1, 2, \ldots, s+1\}$. The solutions of (4-1) form an affine subspace of dimension $k - (k-n) - s - 1 = n - s - 1$, so if $s \le n-2$, then there are at least two solutions, one of which does not satisfy the identity $\iota_k = \sum_{i=1}^{s+1} c_i \omega_i + v$.

Having constructed the sequence $\{\omega_1, \ldots, \omega_{n-1}\}$, in the remaining step the unique solution to (4-1) for $s = n - 1$ completes the orthonormal sequence.    □

## 5. Binary cyclic frames and circulant Gram matrices

Next, we examine a special type of frame whose Gram matrices are circulants. We recall that a cyclic subspace $V$ of $\mathbb{Z}_2^k$ has the property that it is closed under cyclic shifts; that is, the cyclic shift $S$, which is characterized by $Se_j = e_{j+1 \pmod{k}}$, leaves $V$ invariant.

**Definition 5.1.** A frame $\mathcal{F} = \{f_1, f_2, \ldots, f_k\}$ for $\mathbb{Z}_2^n$ is called a binary cyclic frame if the range of the analysis operator is invariant under the cyclic shift $S$. If $\mathcal{F}$ is also Parseval, then we say that is a binary cyclic $(k, n)$-frame.

Since the range of the Gram matrix $G$ belonging to a Parseval frame is identical to the set of eigenvectors corresponding to eigenvalue 1, we have a simple characterization of Gram matrices of binary cyclic Parseval frames.

**Theorem 5.2.** *A binary frame* $\mathcal{F} = \{f_1, f_2, \ldots, f_k\}$ *for* $\mathbb{Z}_2^n$ *is a cyclic Parseval frame if and only if its Gram matrix* $G_{\mathcal{F}}$ *is a symmetric idempotent circulant matrix (that is,* $G_{\mathcal{F}} = G_{\mathcal{F}}^* = G_{\mathcal{F}}^2$ *and* $SG_{\mathcal{F}}S^* = G_{\mathcal{F}}$*), with only odd column vectors.*

*Proof.* If $G_{\mathcal{F}}$ is the Gram matrix of a binary cyclic Parseval frame, then from the Parseval property, we know that $G_{\mathcal{F}} = G_{\mathcal{F}}^* = G_{\mathcal{F}}^2$. Moreover, by the cyclicity of the frame, the eigenspace corresponding to eigenvalue 1 of $G_{\mathcal{F}}$ is invariant under $S$, and thus if $x = G_{\mathcal{F}}x$, then $Sx = SG_{\mathcal{F}}x = G_{\mathcal{F}}Sx$. Using this identity repeatedly and writing $y = S^{k-1}x = S^*x$ gives $y = SG_{\mathcal{F}}S^*y$ for all $y$ in the range of $G_{\mathcal{F}}$. By the symmetry of $G_{\mathcal{F}}$, the range of $G_{\mathcal{F}}$ is identical to that of $G_{\mathcal{F}}^*$, so $\langle G_{\mathcal{F}}x, y \rangle = \langle SG_{\mathcal{F}}S^*x, y \rangle$ for all $x$, $y$ in the range of $G_{\mathcal{F}}$ establishes the circulant property $G_{\mathcal{F}} = SG_{\mathcal{F}}S^*$. If $G_{\mathcal{F}}$ is a circulant, then each column vector generates all the others by applying powers of the cyclic shift to it. Thus, if one column vector is odd, so are all the other column vectors. Applying Theorem 4.1 then yields that the Gram matrices of binary cyclic Parseval frames are symmetric idempotent circulant matrices with only odd column vectors.

| $k$ | first row of matrix | $k$ | first row of matrix | $k$ | first row of matrix |
|---|---|---|---|---|---|
| 3 | 100 | 11 | 10000000000 | 16 | 1000000000000000 |
|   | 111 |    | 11111111111 | 17 | 10000000000000000 |
| 4 | 1000 | 12 | 100000000000 |    | 10010111001110100 |
| 5 | 10000 |   | 100010001000 |    | 11101000110001011 |
|   | 11111 | 13 | 1000000000000 |   | 11111111111111111 |
| 6 | 100000 |   | 1111111111111 | 18 | 100000000000000000 |
|   | 101010 | 14 | 10000000000000 |   | 100000100000100000 |
| 7 | 1000000 |   | 10101010101010 |   | 101010001010001010 |
|   | 1111111 | 15 | 100000000000000 |   | 101010101010101010 |
| 8 | 10000000 |   | 100001000010000 | 19 | 1000000000000000000 |
| 9 | 100000000 |   | 100100100100100 |   | 1111111111111111111 |
|   | 100100100 |   | 100101100110100 | 20 | 10000000000000000000 |
|   | 111011011 |   | 111010011001011 |   | 10001000100010001000 |
|   | 111111111 |   | 111011011011011 |   |   |
| 10 | 1000000000 |  | 111110111101111 |   |   |
|    | 1010101010 |  | 111111111111111 |   |   |

**Table 2.** For $k$ ranging from 3 to 20, the table gives the first row of
the circulant $k \times k$ Gram matrix of each binary cyclic $(k, n)$-frame.

```
1 1 1 0 1 1 0 1 1        1 0 0 1 1 1 1
1 1 1 1 0 1 1 0 1        1 1 1 0 0 1 1
1 1 1 1 1 0 1 1 0        1 1 1 1 1 0 0
0 1 1 1 1 1 0 1 1        0 1 0 1 1 1 1
1 0 1 1 1 1 1 0 1        0 0 0 1 0 1 1
1 1 0 1 1 1 1 1 0        0 0 0 0 0 1 0
0 1 1 0 1 1 1 1 1        0 0 1 1 1 1 1
1 0 1 1 0 1 1 1 1        0 0 0 0 1 1 1
1 1 0 1 1 0 1 1 1        0 0 0 0 0 0 1
```

**Table 3.** Circulant Gram matrix (left) and analysis matrix (right)
of the unique binary cyclic $(9, 7)$-frame with nonrepeating vectors.

Conversely, if $G$ is a symmetric idempotent circulant and each column vector is
odd, then Theorem 4.1 again yields that it is the Gram matrix of a binary Parseval
frame $\mathcal{F}$ with $G = \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*$. Moreover, the range of $G$ is invariant under the cyclic
shift, because one column vector generates all the others by applying powers of the
cyclic shift to it. Since the range of $G$ is identical to that of $\Theta_{\mathcal{F}}$, we have that $\mathcal{F}$ is
a cyclic binary Parseval frame. $\qquad\square$

|  | $n = 7$ |  | $n = 13$ |
| --- | --- | --- | --- |
| 111010011001011 | 1010111 | 111011011011011 | 1001111001111 |
| 111101001100101 | 1101101 | 111101101101101 | 1110011110011 |
| 111110100110010 | 0101100 | 111110110110110 | 1111100111100 |
| 011111010011001 | 1011011 | 011111011011011 | 0101111001111 |
| 101111101001100 | 1101110 | 101111101101101 | 0001011110011 |
| 010111110100110 | 1100100 | 110111110110110 | 0000010111100 |
| 001011111010011 | 0110001 | 011011111011011 | 0011111001111 |
| 100101111101001 | 1000011 | 101101111101101 | 0000111110011 |
| 110010111110100 | 1101000 | 110110111110110 | 0000001111100 |
| 011001011111010 | 0110010 | 011011011111011 | 0000000101111 |
| 001100101111101 | 0001011 | 101101101111101 | 0000000001011 |
| 100110010111110 | 0000010 | 110110110111110 | 0000000000010 |
| 010011001011111 | 0011111 | 011011011011111 | 0000000011111 |
| 101001100101111 | 0000111 | 101101101101111 | 0000000000111 |
| 110100110010111 | 0000001 | 110110110110111 | 0000000000001 |

|  | $n = 9$ |  | $n = 11$ |
| --- | --- | --- | --- |
| 100101100110100 | 101000100 | 111110111101111 | 10011111111 |
| 010010110011010 | 010010010 | 111111011110111 | 11100111111 |
| 001001011001101 | 100000101 | 111111101111011 | 11111001111 |
| 100100101100110 | 111000110 | 111111110111101 | 11111110011 |
| 010010010110011 | 111110011 | 111111111011110 | 11111111100 |
| 101001001011001 | 011011001 | 011111111101111 | 01011111111 |
| 110100100101100 | 101110100 | 101111111110111 | 00010111111 |
| 011010010010110 | 011111110 | 110111111111011 | 00000101111 |
| 001101001001011 | 000111011 | 111011111111101 | 00000001011 |
| 100110100100101 | 000001101 | 111101111111110 | 00000000010 |
| 110011010010010 | 001100010 | 011110111111111 | 00111111111 |
| 011001101001001 | 000011001 | 101111011111111 | 00001111111 |
| 101100110100100 | 000000100 | 110111101111111 | 00000011111 |
| 010110011010010 | 000000010 | 111011110111111 | 00000000111 |
| 001011001101001 | 000000001 | 111101111011111 | 00000000001 |

**Table 4.** Circulant Gram matrix (first matrix of each pair) and analysis matrix (second of pair) of binary cyclic $(15, n)$-frames, $n < k$, whose vectors do not repeat.

Since adding the identity matrix changes odd columns of $G$ to even columns, we conclude that complementary Parseval frames do not exist for binary cyclic Parseval frames.

**Corollary 5.3.** *If $\mathcal{F}$ is a binary cyclic Parseval frame, then it has no complementary Parseval frame.*

In Table 2, we provide an exhaustive list of the Gram matrices of cyclic binary Parseval frames with $3 \leq k \leq 20$. Factoring these into the corresponding analysis

and synthesis matrices shows that many of these examples contain repeated frame vectors. In an earlier paper [Bodmann et al. 2009], such repeated vectors were associated with a trivial form of redundancy incorporated in the analysis matrix $\Theta_{\mathcal{F}}$. Tables 3 and 4 list the circulant Gram matrices of rank $n < k \leq 20$, paired with $k \times n$ analysis matrices, for which no repetition of frame vectors occurs.

# References

[Balan 1999]  R. Balan, "Equivalence relations and distances between Hilbert frames", *Proc. Amer. Math. Soc.* **127**:8 (1999), 2353–2366. MR Zbl

[Betten et al. 2006]  A. Betten, M. Braun, H. Fripertinger, A. Kerber, A. Kohnert, and A. Wassermann, *Error-correcting linear codes*, Algorithms and Computation in Mathematics **18**, Springer, 2006. MR Zbl

[Bodmann and Paulsen 2005]  B. G. Bodmann and V. I. Paulsen, "Frames, graphs and erasures", *Linear Algebra Appl.* **404** (2005), 118–146. MR Zbl

[Bodmann et al. 2009]  B. G. Bodmann, M. Le, L. Reza, M. Tobin, and M. Tomforde, "Frame theory for binary vector spaces", *Involve* **2**:5 (2009), 589–602. MR Zbl

[Bodmann et al. 2014]  B. G. Bodmann, B. Camp, and D. Mahoney, "Binary frames, graphs and erasures", *Involve* **7**:2 (2014), 151–169. MR Zbl

[Christensen 2003]  O. Christensen, *An introduction to frames and Riesz bases*, Birkhäuser, Boston, 2003. MR Zbl

[Duffin and Schaeffer 1952]  R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series", *Trans. Amer. Math. Soc.* **72** (1952), 341–366. MR Zbl

[Goyal et al. 2001]  V. K. Goyal, J. Kovačević, and J. A. Kelner, "Quantized frame expansions with erasures", *Appl. Comput. Harmon. Anal.* **10**:3 (2001), 203–233. MR Zbl

[Haemers et al. 1999]  W. H. Haemers, R. Peeters, and J. M. van Rijckevorsel, "Binary codes of strongly regular graphs", *Des. Codes Cryptogr.* **17**:1 (1999), 187–209. MR Zbl

[Han and Larson 2000]  D. Han and D. R. Larson, *Frames, bases and group representations*, vol. 147, Mem. Amer. Math. Soc. **697**, American Mathematical Society, Providence, RI, 2000. MR Zbl

[Han et al. 2007]  D. Han, K. Kornelson, D. Larson, and E. Weber, *Frames for undergraduates*, Student Mathematical Library **40**, American Mathematical Society, Providence, RI, 2007. MR Zbl

[Holmes and Paulsen 2004]  R. B. Holmes and V. I. Paulsen, "Optimal frames for erasures", *Linear Algebra Appl.* **377** (2004), 31–51. MR Zbl

[Hotovy et al. 2015]  R. Hotovy, D. R. Larson, and S. Scholze, "Binary frames", *Houston J. Math.* **41**:3 (2015), 875–899. MR Zbl

[Kovačević and Chebira 2007a]  J. Kovačević and A. Chebira, "Life beyond bases: the advent of frames, I", *IEEE Signal Process. Mag.* **24**:4 (2007), 86–104.

[Kovačević and Chebira 2007b]  J. Kovačević and A. Chebira, "Life beyond bases: the advent of frames, II", *IEEE Signal Process. Mag.* **24**:5 (2007), 115–125.

[Lempel 1975]  A. Lempel, "Matrix factorization over GF(2) and trace-orthogonal bases of GF($2^n$)", *SIAM J. Comput.* **4** (1975), 175–186. MR Zbl

[MacWilliams and Sloane 1977]  F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes, I*, North-Holland Math. Library **16**, North-Holland Pub., Amsterdam, 1977. MR Zbl

[Marshall 1984]  T. Marshall, "Coding of real-number sequences for error correction: a digital signal processing problem", *IEEE J. Selected Areas Comm.* **2**:2 (1984), 381–392.

[Marshall 1989] T. Marshall, "Fourier transform convolutional error-correcting codes", pp. 653–657 in *Twenty-third Asilomar conference on signals, systems and computers*, edited by R. R. Chen, IEEE, Piscataway, NJ, 1989.

[Puschel and Kovačević 2005] M. Puschel and J. Kovačević, "Real, tight frames with maximal robustness to erasures", pp. 63–72 in *Proceedings DCC 2005: data compression conference*, edited by J. A. Storer and M. Cohn, IEEE, Piscataway, NJ, 2005.

[Rath and Guillemot 2003] G. Rath and C. Guillemot, "Performance analysis and recursive syndrome decoding of DFT codes for bursty erasure recovery", *IEEE Trans. Signal Process.* **51**:5 (2003), 1335–1350. MR

[Rath and Guillemot 2004] G. Rath and C. Guillemot, "Frame-theoretic analysis of DFT codes with erasures", *IEEE Trans. Signal Process.* **52**:2 (2004), 447–460. MR

zacherybaker96@gmail.com        *Department of Mathematics, University of Houston, Houston, TX, United States*

bgb@math.uh.edu        *Department of Mathematics, University of Houston, Houston, TX, United States*

micahbullock@outlook.com        *Department of Mathematics, University of Houston, Houston, TX, United States*

sambranum@gmail.com        *Department of Mathematics, University of Houston, Houston, TX, United States*

mclaneyjacob@gmail.com        *Department of Mathematics, University of Houston, Houston, TX, United States*

# Numbers and the heights of their happiness

## May Mei and Andrew Read-McFarland

(Communicated by Kenneth S. Berenhaut)

A generalized happy function, $S_{e,b}$ maps a positive integer to the sum of its base $b$ digits raised to the $e$-th power. We say that $x$ is a base-$b$, $e$-power, height-$h$, $u$-attracted number if $h$ is the smallest positive integer such that $S_{e,b}^h(x) = u$. Happy numbers are then base-10, 2-power, 1-attracted numbers of any height. Let $\sigma_{h,e,b}(u)$ denote the smallest height-$h$, $u$-attracted number for a fixed base $b$ and exponent $e$ and let $g(e)$ denote the smallest number such that every integer can be written as $x_1^e + x_2^e + \cdots + x_{g(e)}^e$ for some nonnegative integers $x_1, x_2, \ldots, x_{g(e)}$. We prove that if $p_{e,b}$ is the smallest nonnegative integer such that $b^{p_{e,b}} > g(e)$,

$$d = \left\lceil \frac{g(e) + 1}{1 - \left(\frac{b-2}{b-1}\right)^e} + e + p_{e,b} \right\rceil,$$

and $\sigma_{h,e,b}(u) \geq b^d$, then $S_{e,b}(\sigma_{h+1,e,b}(u)) = \sigma_{h,e,b}(u)$.

## 1. Introduction

Let $S_{e,b}$ be the function that maps a positive base-$b$ integer to the sum of its digits raised to the $e$-th power, where $e$ is a positive integer. That is, for $x = \sum_{i=0}^{n-1} a_i b^i$, with $0 \leq a_i \leq b - 1$ for all $i$,

$$S_{e,b}\left(\sum_{i=0}^{n-1} a_i b^i\right) = \sum_{i=0}^{n-1} a_i^e.$$

If $S_{e,b}^h(x) = 1$ for some integer $h$, then $x$ is said to be an $e$-power, $b$-happy number. Guy [2004] gave the smallest 2-power, 10-happy numbers of heights 0 through 6 and asked if 78999 is the smallest height-7 happy number. Grundman and Teeple [2003] answered Guy, giving the smallest 2-power, 10-happy numbers of heights 0 through 10, and 3-power, 10-happy numbers of heights 0 through 8. From Grundman and Teeple's work, one can extract an algorithm for finding the smallest happy number of height $h + 1$ if the smallest happy number of height $h$ is known. The main results of this paper are Theorems 3.1 and 3.3, which jointly imply that once

the smallest height-$(h+1)$, $u$-attracted, base-$b$ number is sufficiently large, applying $S_{e,b}$ to that number will yield the smallest height-$h$, $u$-attracted, base-$b$ number. The results of this paper hold not only for happy numbers (i.e., 1-attracted), but more generally for $u$-attracted numbers. Moreover, our results hold for all bases and exponents.

**Definition 1.1.** For a fixed base $b$, exponent $e$, and positive integer $u$, we say that a positive integer $x$ is $u$-attracted if $S_{e,b}^n(x) = u$ for some nonnegative integer $n$. If $h$ is the smallest nonnegative integer so that $S_{e,b}^h(x) = u$ then $x$ is a height-$h$, $u$-attracted number. (As a convention, $S_{e,b}^0(x) = x$.)

**Definition 1.2.** For a fixed base $b$, exponent $e$, positive integer $u$, and nonnegative integer $h$, let $\sigma_{h,e,b}(u)$ denote the smallest height-$h$, $u$-attracted number, that is, the smallest positive integer $k$ with the property that $S_{e,b}^h(k) = u$ and $S_{e,b}^n(k) \neq u$ for $n < h$. Similarly, for positive $h$, let $\tau_{h,e,b}(u)$ denote the second smallest height-$h$, $u$-attracted number, that is, $S_{e,b}^h(l) = u$, $S_{e,b}^n(l) \neq u$ for $n < h$, and $\sigma_{h,e,b}(u) < l$.

Some of the following proofs rely upon knowing the smallest integer $x$ such that for a given $e$, every integer is expressible as the sum of at most $x$ many integers raised to the $e$-th power. We define $g(e)$ for this purpose.

**Definition 1.3.** For a fixed positive integer $e$, let $g(e)$ denote the smallest integer such that every nonnegative integer is expressible as $x_1^e + x_2^e + \cdots + x_{g(e)}^e$, where $x_1, x_2, \ldots, x_{g(e)}$ are all nonnegative integers.

This is the well-known Waring's problem. Many surveys about the history of this problem exist; see for instance [Vaughan and Wooley 2002].

For the entirety of this paper, we assume that the base $b \geq 2$ is an integer, the exponent $e \geq 1$ is an integer, the height $h$ is a nonnegative integer, the attractor $u$ is a positive integer, and that $x$ denotes a positive integer. Additionally, when we say $\lceil x \rceil = y$ we mean that $y$ is the smallest integer such that $y \geq x$, and similarly, if $\lfloor x \rfloor = y$, then $y$ is the largest integer such that $y \leq x$.

## 2. Mapping attracted numbers

In this section, we establish in Theorem 2.2 a criterion, depending on $g(e)$ that ensures that $S_{e,b}(\sigma_{h+1,e,b}(u)) = \sigma_{h,e,b}(u)$ for a fixed base $b$, exponent $e$, height $h$, and integer $u$.

**Lemma 2.1.** *Fix a base $b$, exponent $e$, and attractor $u$. The smallest positive integer $x$ such that $S_{e,b}(x) = u$ has $n$ digits, where*

$$\frac{u}{(b-1)^e} \leq n \leq \frac{u}{(b-1)^e} + g(e).$$

*Proof.* Since the maximum value of the image of each digit under $S_{e,b}$ is $(b-1)^e$, $u/(b-1)^e$ is a lower bound for the number of digits of $x$. Let $q$ and $r$ be the quotient and remainder of $u$ divided by $(b-1)^e$, respectively; that is, $q$ is a nonnegative integer, $0 \le r < (b-1)^e$, and $u = q(b-1)^e + r$. Let $x_1, \ldots, x_{g(e)}$ be integers such that $x_1^e + \cdots + x_{g(e)}^e = r$. Since $r < (b-1)^e$, we have $x_1, \ldots, x_{g(e)} < b-1$ and so they are valid digits in base $b$. Without loss of generality, $x_1 \le x_2 \le \cdots \le x_{g(e)}$. Let $y$ be the positive integer formed by the digits $x_1, x_2, \ldots, x_{g(e)}$ followed by $q$ digits, each of which is $b-1$. Since $x$ is minimal, it follows that $x \le y$. So $n$, the number of digits of $x$, must be less than or equal to the number of digits of $y$, which is $\lfloor u/(b-1)^e \rfloor + g(e)$. $\qquad\square$

**Theorem 2.2.** *Fix a base $b$, exponent $e$, positive height $h$, and attractor $u$. If*

$$\frac{\sigma_{h,e,b}(u)}{(b-1)^e} + g(e) \le \frac{\tau_{h,e,b}(u)}{(b-1)^e}, \tag{1}$$

*then $S_{e,b}(\sigma_{h+1,e,b}(u)) = \sigma_{h,e,b}(u)$.*

*Proof.* Let $x$ be the smallest integer such that $S_{e,b}(x) = \sigma_{h,e,b}(u)$. Let $z$ be a height-$h$, $u$-attracted number that is greater than $\sigma_{h,e,b}(u)$ (recall that $\tau_{h,e,b}$ is the smallest such number) and $y$ any integer such that $S_{e,b}(y) = z$. That is, $y$ is a height-$(h+1)$, $u$-attracted number whose image is not $\sigma_{h,e,b}(u)$. Let $n$ be the number of digits of $x$ and $m$ be the number of digits of $y$. We will show that $x < y$. By Lemma 2.1,

$$n \le \frac{\sigma_{h,e,b}(u)}{(b-1)^e} + g(e) \quad \text{and} \quad \frac{\tau_{h,e,b}(u)}{(b-1)^e} \le \frac{z}{(b-1)^e} \le m.$$

By the hypothesis (1), this gives $n \le m$. If $n < m$, then $x < y$, so let us suppose that $n = m$. It must then be the case that

$$\frac{\sigma_{h,e,b}(u)}{(b-1)^e} + g(e) = \frac{z}{(b-1)^e}.$$

Since $S_{e,b}(y) = z$ and $y$ has $m = z/(b-1)^e$ digits, $y$ is the concatenation of $m$ digits, each of which is $b-1$. Since $x \ne y$ (as they have different images under $S_{e,b}$) and $x$ and $y$ have the same number of digits, at least one digit of $x$ is not $b-1$. Thus, $x < y$. Hence $x$ is less than every other height-$(h+1)$, $u$-attracted number, and so $x = \sigma_{h+1,e,b}(u)$. Since $S_{e,b}(x) = \sigma_{h,e,b}(u)$, we have $S_{e,b}(\sigma_{h+1,e,b}(u)) = \sigma_{h,e,b}(u)$. $\qquad\square$

From [Grundman and Teeple 2003], it is known that $\sigma_{7,2,10} = 78999$ and $\tau_{7,2,10}(1) = 79899$.

**Question 2.3.** *Under what conditions is $\tau_{h,e,b}(u)$ a permutation of the digits of $\sigma_{h,e,b}(u)$?*

## 3. Large $u$-attracted numbers

In this section, we prove Theorems 3.1 and 3.3, which imply that once $\sigma_{h,e,b}(u)$ is sufficiently large, $S_{e,b}(\sigma_{h+1,e,b}(u)) = \sigma_{h,e,b}(u)$.

**Theorem 3.1.** *Fix a base $b$, exponent $e$, positive height $h$, and attractor $u$. Let $\delta$ be a positive integer, and let*

$$d = \frac{g(e) + 1}{1 - \left(\frac{b-2}{b-1}\right)^e} + \delta.$$

*If $\sigma_{h,e,b}(u)$ has at least $d$ digits, then the base-$b$ expansion of $\sigma_{h,e,b}$ is of the form*

$$\sigma_{h,e,b}(u) = \sum_{i=0}^{n-1} a_i b^i$$

*with $a_0, \ldots, a_\delta = b - 1$. More informally, the rightmost $\delta + 1$ digits of $\sigma_{h,e,b}(u)$ are all $b - 1$.*

*Proof.* In this proof, we will show that if $\sigma_{h,e,b}$ has "too many" digits which are not equal to $b - 1$, we can construct a smaller number with the same image as $\sigma_{h,e,b}$. This contradicts the definition of $\sigma_{h,e,b}$.

One can verify $\sigma_{1,e,b}(1) = 10$ (in base $b$) for all $e$, $b$ and that this is the only number of the form $\sigma_{h,e,b}$ with a 0 digit. However, 10 is a two-digit number and $d > 2$ for integers $e > 1$. Thus, using the base-$b$ expansion from the statement of the theorem, $a_{i+1} \le a_i$ for $0 \le i < n - 1$ (its digits must appear in increasing order from left to right) and none of its digits can be 0 since $\sigma_{h,e,b}(u)$ is the least height-$h$, $u$-attracted number.

In the case $a_i = b - 1$ for all $i$, this theorem is trivially true. Otherwise, let us construct $z$, the sum of the image of the digits which are not equal to $b - 1$. In the case that some digits of $\sigma_{h,e,b,}(u)$ are $b - 1$ and some are not, define an integer parameter $k \ge 2$ to be such that $a_{k-1} < b - 1$ and for all $i < k - 1$, $a_i = b - 1$. That is, the $k$-th place is the first (from the right) in which a digit that is not $b - 1$ appears. Hence,

$$\sigma_{h,e,b}(u) = \sum_{i=k-1}^{n-1} a_i b^i + \sum_{i=0}^{k-2} (b-1) b^i.$$

Let $y = S_{e,b}(\sigma_{h,e,b}(u))$ and let $z = y - (k-1)(b-1)^e$, that is,

$$z = \sum_{i=k-1}^{n-1} a_i^e.$$

In the case that no digits of $\sigma_{h,e,b}$ are $b - 1$, set $k = 1$ and let $z = \sum_{i=0}^{n-1} a_i^e$. We proceed to show that if $k \le \delta + 1$, we can construct a number smaller than $\sigma_{h,e,b}$ with the same image as $\sigma_{h,e,b}$, a contradiction. Let $n' = n - (k-1)$ and

let $m = \lfloor z/(b-1)^e \rfloor$. Since $z$ is the sum of $n'$ many terms of the form $a_i^e$, where $a_i \le b-2$ for all $i$, we have $n' \ge z/(b-2)^e$. Thus,

$$\frac{(b-2)^e}{(b-1)^e} n' \ge \frac{z}{(b-1)^e} \ge m.$$

So,

$$\left(\frac{b-2}{b-1}\right)^e n' + g(e) + 1 \ge m + g(e) + 1.$$

By the definition of $d$,

$$d - \delta = \frac{g(e)+1}{1 - \left(\frac{b-2}{b-1}\right)^e},$$

and since $k \le \delta + 1$,

$$d - (k-1) \ge \frac{g(e)+1}{1 - \left(\frac{b-2}{b-1}\right)^e}.$$

Thus,

$$(d - (k-1))\left(1 - \left(\frac{b-2}{b-1}\right)^e\right) \ge g(e) + 1.$$

And since $n' \ge d - (k-1)$ and $1 - \left(\frac{b-2}{b-1}\right)^e > 0$, we have

$$n'\left(1 - \left(\frac{b-2}{b-1}\right)^e\right) \ge g(e) + 1$$

and hence

$$n' \ge g(e) + 1 + n'\left(\frac{b-2}{b-1}\right)^e \ge m + g(e) + 1.$$

Therefore, $n' \ge m + g(e) + 1$.

Let $r$ be the remainder of $y$ divided by $(b-1)^e$; that is, $y = q(b-1)^e + r$, where $q \ge 0$ and $(b-1)^e > r \ge 0$. From the definition of $m$, we have $q = m + (k-1)$. Let $x_1, x_2, \ldots, x_{g(e)}$ be integers less than $b-1$ so that $x_1^e + x_2^e + \cdots + x_{g(e)}^e = r$. There are such $x_j$ since $g(e)$ is defined so that such integers exist, and all integers must be less than $b-1$ since $r < (b-1)^e$. Without loss of generality, $x_1 \le x_2 \le \cdots \le x_{g(e)}$. Let $x$ be a base-$b$ number with digits $x_1, \ldots, x_{g(e)}$ followed by $m + (k-1)$ many $b-1$ digits.

Hence, $S_{e,b}(x) = y$, and $x$ has at most $g(e) + m + (k-1)$ digits. Since $n' = n - (k-1)$, we know $n \ge g(e) + 1 + m + (k-1)$. However, this means that $x$ has fewer digits than $\sigma_{h,e,b}(u)$. This contradicts the fact that $\sigma_{h,e,b}(u)$ is the smallest height-$h$, $u$-attracted integer, and hence, $k > \delta + 1$. $\qquad\square$

For ease of notation, we define a constant $p_{e,b}$.

**Definition 3.2.** For a fixed exponent $e$ and base $b$, let $p_{e,b}$ be the smallest integer such that $b^{p_{e,b}} > g(e)$.

**Theorem 3.3.** *Fix a base $b$, exponent $e$, positive height $h$, and attractor $u$. If $\sigma_{h,e,b}(u) = \sum_{i=0}^{n-1} a_i b^i$, where $a_0, \ldots, a_{e+p_{e,b}} = b - 1$, then $S_{e,b}(\sigma_{h+1,e,b}(u)) = \sigma_{h,e,b}(u)$.*

*Proof.* Let $\sigma_{h,e,b}(u)$ be such that $a_0, \ldots, a_k = b - 1$, where $k \geq e + p_{e,b}$. Define $c_j = \sigma_{h,e,b}(u) + j$ for $1 \leq j < g(e)(b-1)^e$. We will show that $c_1$ through $c_{g(e)(b-1)^e - 1}$ are not height-$h$, $u$-attracted numbers.

If $b > 2$, using the definition of $p_{e,b}$ we get

$$j < g(e)(b - 1)^e < b^{p_{e,b}}(b - 1)^e < b^{p_{e,b}} b^e = b^{e+p_{e,b}}.$$

Since $\sigma_{h,e,b}$ has at least $e + p_{e,b} + 1$ trailing digits equal to $b - 1$, we know $c_1$ has at least $e + p_{e,b} + 1$ trailing zeros. Since $j < b^{e+p_{e,b}}$, we know $j$ has at most $e + p_{e,b}$ many digits. Hence $c_j$ has at least one digit which is zero for $1 \leq j < g(e)(b-1)^e$. Let $c_j'$ be formed by removing the all zero digits of $c_j$. We claim that $c_j' < \sigma_{h,e,b}(u)$. Recall that $n$ denotes the number of digits of $\sigma_{h,e,b}(u)$. If $a_i \neq b-1$ for some $i$, then $n \geq e + p_{e,b} + 2$ and $c_j$ has $n$ digits for all $j$. Thus, $c_j'$ has at most $n-1$ digits and hence $c_j' < \sigma_{h,e,b}$. If $a_i = b - 1$ for all $i$, then $\sigma_{h,e,b}(u) = b^n - 1$ and $c_1 = b^n = b^{e+p_{e,b}+1}$, which means that $c_j < b^{e+p_{e,b}+1} + b^{e+p_{e,b}}$. Thus $c_j'$ has at most $n$ digits, while the leading digit of $\sigma_{h,e,b}$ is $b-1$, but the leading digit of $c_j'$ is 1, and since $b \neq 2$, $c_j' < \sigma_{h,e,b}$.

This leaves only the case that $b = 2$. In this case,

$$j < g(e)(2 - 1)^e = g(e) < 2^{p_{e,2}}.$$

Since the only allowable digits are 0 and 1, and we argued in the proof of Theorem 3.1 that $\sigma_{h,e,b}$ does not have any digits that are equal to zero, $\sigma_{h,e,2} = 2^{n+1} - 1$ for some $n \geq e + p_{e,2}$, so $2^{n+1} \leq c_j < 2^{n+1} + 2^{p_{e,2}}$ for all $j$. Since $n \geq e + p_{e,2}$ and $e$ is at least 1, $c_j$ has at least two digits that are equal to 0. Again, let $c_j'$ be formed by removing the all zero digits of $c_j$. Then $c_j'$ has fewer than $n$ digits and hence $c_j' < \sigma_{h,e,2}$.

So, if any $c_j$ are height-$h$, $u$-attracted numbers, then $c_j'$ is a smaller height-$h$, $u$-attracted number than $\sigma_{h,e,b}(u)$, contradicting the definition of $\sigma_{h,e,b}(u)$. Hence, $\tau_{h,e,b}(u) \geq g(e)(b-1)^e + \sigma_{h,e,b}(u)$. Therefore, by Theorem 2.2, $S_{e,b}(\sigma_{h+1,e,b}) = \sigma_{h,e,b}$. $\square$

**Corollary 3.4.** *Fix a base $b$ and exponent $e$. Let*

$$d = \left\lceil \frac{g(e) + 1}{1 - \left(\frac{b-2}{b-1}\right)^e} + e + p_{e,b} \right\rceil.$$

*If $\sigma_{h,e,b}(u) \geq b^d$, then $S_{e,b}(\sigma_{h+1,e,b}(u)) = \sigma_{h,e,b}(u)$.*

*Proof.* Since $\sigma_{h,e,b}(u) \geq b^d$, we know $\sigma_{h,e,b}(u)$ must have at least $d - 1$ digits. Hence, by Theorem 3.1, $\sigma_{h,e,b}(u) = \sum_{i=0}^{n-1} a_i b^i$, where for $i \leq e + p_{e,b}$, we have $a_i = b - 1$. Therefore, by Theorem 3.3, $S_{e,b}(\sigma_{h+1,e,b}(u)) = \sigma_{h,e,b}(u)$. $\square$

Corollary 3.4 gives a bound $b^d$ for $\sigma_{h,e,b}(u)$ (in terms of $e$ and $b$) so that if $\sigma_{h,e,b}(u) \geq b^d$, then $S_{e,b}(\sigma_{h+1,e,b}(u)) = \sigma_{h,e,b}$. This leads to the natural question:

**Question 3.5.** *Is there a bound $\beta$ for $h$ (in terms of $e$ and $b$) so that if $h \geq \beta$, $S_{e,b}(\sigma_{h+1,e,b}(u)) = \sigma_{h,e,b}$?*

## Acknowledgements

## References

[Grundman and Teeple 2003] H. G. Grundman and E. A. Teeple, "Heights of happy numbers and cubic happy numbers", *Fibonacci Quart.* **41**:4 (2003), 301–306. MR Zbl

[Guy 2004] R. K. Guy, *Unsolved problems in number theory*, 3rd ed., Springer, 2004. MR Zbl

[Vaughan and Wooley 2002] R. C. Vaughan and T. D. Wooley, "Waring's problem: a survey", pp. 301–340 in *Number theory for the millennium, III* (Urbana, IL, 2000), edited by M. A. Bennett et al., A K Peters, Natick, MA, 2002. MR Zbl

meim@denison.edu                     *Department of Mathematics & Computer Science, Denison University, Granville, OH, United States*

readmc_a1@denison.edu                     *Department of Mathematics & Computer Science, Denison University, Granville, OH, United States*

msp

# The truncated and supplemented Pascal matrix and applications

Michael Hua, Steven B. Damelin, Jeffrey Sun and Mingchao Yu

(Communicated by Jim Haglund)

In this paper, we introduce the $k \times n$ (with $k \leq n$) truncated, supplemented Pascal matrix, which has the property that any $k$ columns form a linearly independent set. This property is also present in Reed–Solomon codes; however, Reed–Solomon codes are completely dense, whereas the truncated, supplemented Pascal matrix has multiple zeros. If the maximum distance separable code conjecture is correct, then our matrix has the maximal number of columns (with the aforementioned property) that the conjecture allows. This matrix has applications in coding, network coding, and matroid theory.

## 1. Introduction

Finite field linear algebra is an import branch of linear algebra. Instead of using the infinite field $\mathbb{R}$, it uses linearly independent vectors consisting of a finite number of elements, which can be represented by a finite number of bits. It has thus motivated many practical coding techniques, such as Reed–Solomon codes [1960] and linear network coding [Li et al. 2003; Ho et al. 2006]. It is also closely related to structural matroid theory through matroid representability [Oxley 2011; Oxley et al. 1996; El Rouayheb et al. 2010; Yu et al. 2014].

One of the most important problems in finite field linear algebra is finding the size of the largest set of vectors over a $k$-dimensional finite field such that every subset of $k$ vectors is linearly independent [Ball 2012; Ball and De Beule 2012]. From a matrix perspective, the problem is described as:

**Problem 1.1.** *Consider a finite field $\mathbb{F}_q$, where $q = p^h$, for $p$ a prime and $h$ a nonnegative integer. Given a positive integer $k$, what is the largest integer $n$ such that there exists a $k \times n$ matrix $\mathbf{H}$ over $\mathbb{F}_q$, in which every set of $k$ columns is linearly independent?*

Such a matrix, upon its existence, could be the generator matrix of an $[n, k]$ maximum distance separable (MDS) code [Lin and Costello 2004], which can correct up to $d = n - k$ bits of erasures or $t = d/2$ bits of errors. We will thus refer to $H$ as an MDS matrix. Its existence also determines the representability of uniform matroids, which we will discuss in detail in Section 4C. The maximal value of $n$, according to the MDS conjecture, is $q + 1$, unless $q = 2^h$ and $k = 3$ or $k = q - 1$, in which case $n \leq q + 2$. This conjecture has been recently proved for any $q = p$ in [Ball 2012; Ball and De Beule 2012], but a complete proof of it remains open.

Therefore, it is crucial to understand the construction of $k \times (q+1)$ MDS matrices. In coding theory literature, many construction algorithms have been proposed to meet certain coding requirements. However, their computational complexity is not necessarily satisfactory. On one hand, multiplications and additions over large finite fields are required in the matrix construction. On the other hand, the resultant MDS matrix may have low sparsity (or high density), which is measured by the number of zeros in the matrix. Low sparsity can be translated into higher encoding and decoding complexity. It is an open question how these algorithms can be generalized and provide new insights into related fields, such as network coding theory and matroid theory.

In this paper, we investigate the above problems by first proposing in Section 2 a new type of MDS matrix called a *supplemented Pascal matrix*. A supplemented Pascal matrix can be generated by additions and, in particular, without multiplications. It also has guaranteed number of zero entries for high sparsity. We will prove that a supplemented Pascal matrix is an MDS matrix in Section 3. We will then extend our results into a general code construction framework in Section 4A, and then discuss its applications to network coding theory and matroid theory in Sections 4B and 4C, respectively.

## 2. Definitions

For clarity we should first label the elements of a finite field. Henceforth, let $p$ be a prime and $h$ be a nonnegative integer. A finite field $\mathbb{F}_q$ contains $q = p^h$ elements, each represented by a polynomial $g(x) = \sum_{i=0}^{h-1} \beta_i x^i$, whose coefficients are $\{\beta_i\}_{i=0}^{h-1} \in [0, p-1]$. The elements $g(x)$ take on distinct values between 0 and $q - 1$ at $x = p$, which can be used as an intuitive index of the elements. Specifically, we define a index function $\sigma_q(n)$:

**Definition 2.1.** For any integer $n \in [0, q - 1]$, $\sigma_q(n)$ is the element of $\mathbb{F}_q$ whose polynomial coefficients satisfy $\sum_{i=0}^{h-1} \beta_i p^i = n$.

For example, given $q = 2^3$, we have $\sigma_q(0) = 0$, $\sigma_q(1) = 1$, and $\sigma_q(5) = x^2 + 1$.

Based on $\sigma_q(n)$, we define a finite field binomial polynomial $f_{m,q}(n)$:

$$f_{m,q}(n) = \begin{cases} 1 = [\sigma_q(n)]^m, & m = 0, \\ \displaystyle\prod_{i=1}^{m} \frac{\sigma_q(n) - \sigma_q(i-1)}{\sigma_q(i)}, & m > 0, \end{cases} \tag{1}$$

where $\{m, n\} \in [0, q-1]$ are nonnegative integers. Intuitively, $f_{m,q}(n)$ is a polynomial in $\sigma_q(n)$ of degree $m$.

Based on $f_{m,q}(n)$, we introduce the key matrix in this paper, called the *Pascal matrix*:

**Definition 2.2.** Define the matrix $P_q$ over $\mathbb{F}_q$ as the $q \times q$ matrix with elements $P_q(m, n) = f_{m,q}(n)$:

$$P_q = \begin{bmatrix} f_{0,q}(0) & f_{0,q}(1) & \cdots & f_{0,q}(q-1) \\ f_{1,q}(0) & f_{1,q}(1) & \cdots & f_{1,q}(q-1) \\ \vdots & \vdots & \ddots & \vdots \\ f_{q-1,q}(0) & f_{q-1,q}(1) & \cdots & f_{q-1,q}(q-1) \end{bmatrix}. \tag{2}$$

Note that $f_{m,q}(n) = 0$ for $m > n$ and so $P_q$ is an upper-triangular Pascal matrix. For brevity, we simply call it the Pascal matrix.

Note that the matrix index starts from 0.

**Example 2.3.** When $q = 2^2 = 4$, we have

$$P_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & x & x+1 \\ 0 & 0 & 1 & x+1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Our considered matrix $P_q$ is named after Pascal because its entries are binomial coefficients, which is the same as the traditional Pascal matrix, except that the field applied here is $\mathbb{F}_q$, as opposed to $\mathbb{Z}_{\geq 0}$ in the traditional case. Indeed, when $q = p$, the matrix $P_p$ is equal to the traditional Pascal matrix modulo $p$.

**Example 2.4.** When $q = p = 5$, the traditional Pascal matrix $P_{5,T}$ and our Pascal matrix $P_5$ are given by

$$P_{5,T} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 3 & 6 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad P_5 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Indeed, the construction of the Pascal matrix over $\mathbb{F}_p$ shares the same additive formula as the traditional Pascal matrix. Explicitly, $P_p(m, n) = P_p(m-1, m-1) + P_p(m, n-1)$ for every pair of $\{m, n\} \in [1, q-1]$ (note that addition is modulo $p$). This idea appears in Section 4B.

**Definition 2.5.** The truncated Pascal matrix $P_{q,k}$ is the Pascal matrix $P_q$ truncated to the first $k$ rows.

**Example 2.6.** Consider the matrix $P_5$ given in Example 2.4. The truncated Pascal matrix $P_{5,2}$ is given by

$$P_{5,2} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix}.$$

**Definition 2.7.** A *supplemented Pascal matrix*, denoted by $H_{q,k}$, is a truncated Pascal matrix $P_{q,k}$ appended with a column vector $s_k$, which has a one in the bottom entry and zeros everywhere else:

$$H_{q,k} = \left[\; U_{q,k} \;\left|\; \begin{matrix} 0 \\ \vdots \\ 0 \\ 1 \end{matrix} \right.\right]. \tag{3}$$

**Example 2.8.** Supplementing the matrix $P_{5,2}$ in Example 2.6 gives

$$H_{5,2} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & 3 & 4 & 1 \end{bmatrix}.$$

Our supplemented Pascal matrix has a desirable property, namely:

**Theorem 2.9.** *Any k columns of $H_{q,k}$ are linearly independent.*

## 3. Proof of Theorem 2.9

We will first prove the following property of $P_{q,k}$, and then prove that $H_{q,k}$ preserves this property.

**Lemma 3.1** (truncation lemma). *Any k columns of $P_{q,k}$ are linearly independent.*

*Proof.* We first note that $P_q$ (and thus $P_{q,k}$) has the important property that all the entries in the $m$-th row (recall $m$ begins at 0) are defined by the same polynomial $f_{m,q}(n)$, which is a polynomial in $\sigma_q(n)$ of degree $m$. Recall also that $P_q$ (and thus $P_{q,k}$) is upper-triangular. Indeed, we have that $f_{m,q}(n)$ has $m$ roots in $\mathbb{F}_q$ (counting multiplicity). Consequently, the first $m$ entries of the $m$-th row are all zeros.

Given a truncated Pascal matrix $P_{q,k}$, suppose there exist $k$ distinct values of $n$ such that the columns $\{n_0, n_1, \ldots, n_{k-1}\}$ of $P_{q,k}$ constitute a linearly dependent

set. In other words, there exists a $k \times k$ submatrix $M$ of $P_{q,k}$,

$$
M = \begin{bmatrix}
f_{0,q}(n_0) & f_{1,q}(n_1) & \cdots & f_{1,q}(n_{k-1}) \\
f_{1,q}(n_0) & f_{2,q}(n_1) & \cdots & f_{2,q}(n_{k-1}) \\
\vdots & \vdots & \ddots & \vdots \\
f_{k-1,q}(n_0) & f_{k-1,q}(n_1) & \cdots & f_{k-1,q}(n_{k-1})
\end{bmatrix},
\tag{4}
$$

whose rank is smaller than $k$.

If this is the case, then there must exist a nonzero vector $[a_0, a_1, \ldots, a_{k-1}] \in \mathbb{F}_q$ such that $aM = z$, where $z = [z_0, z_1, \ldots, z_{k-1}] = [0, 0, \ldots, 0]$:

$$
[a_0, a_1, \ldots, a_{k-1}]M = [0, 0, \ldots, 0].
$$

Recall that the $m$-th row of $P_{q,k}$ (and thus $M$) is defined by $f_{m,q}(n)$. Correspondingly, $z$ is defined by

$$
f'_q(n) \triangleq \sum_{m=0}^{k-1} \alpha_m f_{m,q}(n),
$$

where $0 = z(i) = f'_q(n_i) = 0$ for all $i \in [0, k-1]$. We also note that the degree of $f'_q(n)$ is at most $k - 1$, because the highest degree of its summands is the degree of $f_{k-1,q}(n)$ with a value of $k - 1$.

Therefore if the columns $\{n_0, n_1, \ldots, n_{k-1}\}$ of $P_{q,k}$ constitute a linearly dependent set, then we will obtain a polynomial $f'_q(n)$ such that

- its degree is at most $k - 1$;
- it has $k$ roots, whose values are $\{\sigma_q(n_0), \sigma_q(n_1), \ldots, \sigma_q(n_{k-1})\}$.

However, with a degree of at most $k - 1$, $f'_q(n)$ can only have at most $k - 1$ roots unless $f'_q(n) = 0$, which is not the case because $a$ is nonzero. Hence, $f'_q(n)$ does not exist, and thus our hypothesis is invalid. Therefore, every $k$ columns of $P_{q,k}$ are linearly independent. Thus Lemma 3.1 is proved.  □

Since $H_{q,k}$ is constructed by appending $s_k$ to $P_{q,k}$, to prove Theorem 2.9 we only need to prove that any $k - 1$ columns of $P_{q,k}$ and $s_k$ together never constitute a linearly dependent set. To see this, we can simply use $s_k$ to linearly cancel the first $q$ entries in the last row of $H_{q,k}$. This will transform $H_{q,k}$ from (3) into

$$
H'_{q,k} = \begin{bmatrix}
 & & & 0 \\
 & P_{q,k-1} & & 0 \\
 & & & \vdots \\
0 & \cdots & 0 & 1
\end{bmatrix},
\tag{5}
$$

which indicates that $s_k$ is orthogonal to all the other columns of $H'_{q,k}$. Then, by applying the truncation lemma to $P_{q,k-1}$, we know that every $k - 1$ out of the first $q$ columns of $H'_{q,k}$ are linearly independent. Adding $s_k$ to them will yield a linearly independent set of $k$. Theorem 2.9 is thus proved.

## 4. Applications

**4A.** *Coding theory.* The truncation lemma can be immediately generalized to any appropriately defined $k \times n$ matrix over $\mathbb{F}_q$ that satisfies (1) $n \le q$, and (2) the $m$-th row ($m \in [0, k-1]$) is defined by a polynomial of degree $m$. For example, by setting $f_{m,q}(n) = \sigma_q(n)^{m-1}$, we can obtain a $k \times n$ matrix over $\mathbb{F}_q$ such that every set of $k$ columns is a linearly independent set. Indeed, this matrix is the generator matrix $G$ of an $[n, k]$ Reed–Solomon code:

$$\begin{bmatrix} \sigma_q(1)^0 & \sigma_q(2)^0 & \cdots & \sigma_q(n)^0 \\ \sigma_q(1) & \sigma_q(2) & \cdots & \sigma_q(n) \\ \sigma_q(1)^2 & \sigma_q(2)^2 & \cdots & \sigma_q(n)^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_q(1)^{k-1} & \sigma_q(2)^{k-1} & \cdots & \sigma_q(n)^{k-1} \end{bmatrix}.$$

Then by appending $s_k$, we can obtain an $[n+1, k]$ extended Reed–Solomon code. Therefore, our polynomial approach is a general approach of constructing *nontrivial* $[n, k]$ MDS codes. It also indicates that the maximum length of any MDS code is at least $q+1$ for any $k \le q$. This result is in agreement with the MDS conjecture [Ball 2012; Ball and De Beule 2012].

Among all the possible constructions, the supplemented Pascal matrix $H_{q,k}$ enjoys a high sparsity, which is the number of zeros in the matrix. Higher sparsity is advantageous, because it generally leads to easier encoding/decoding. However, the sparsity has an upper bound. In the following lemma, we will prove that $H_{q,k}$ approximates this bound with a factor of $\frac{1}{2}$:

**Lemma 4.1** (matrix sparsity). *The number of zeros in the supplemented Pascal matrix $H_{q,k}$ is $\frac{1}{2}$ of the maximum sparsity of any $[n, k]$ code.*

*Proof.* Since any $k \times k$ submatrix of $G$ has a rank of $k$, there is no all-zero row in this matrix. Hence, there are at most $k-1$ zeros in each row of $G$, and at most $k^2 - k$ zeros in total. Recall that in $H_{q,k}$ the $m$-th row has $m+1$ zeros for $m \in [0, k-2]$ and the $(k-1)$-th row has $k-1$ zeros. The total number of zeros is $\frac{1}{2}(k^2 + k - 2)$.     □

**4B.** *Network coding theory.* Network coding (NC) is a class of packet-based coding techniques. Consider a block of $K \ge 1$ data packets $\{x_k\}_{k=0}^{K-1}$, each containing $L$ bits of information. NC treats these data packets as $K$ variables, and sends in the $u$-th ($u \in \mathbb{Z}_{\ge 0}$) transmission a linear combination $y_u$ of all of them:

$$y_u = \sum_{k=0}^{K-1} \alpha_{k,u} x_k, \tag{6}$$

where coefficients $\{\alpha_{k,u}\}_{k=0}^{K-1}$ are elements of a finite field $\mathbb{F}_q$.

Ideally, NC is able to allow any receiver that has received any $K$ coded packets to decode all the $K$ data packets by solving a set of $K$ linear equations. To this end, the associated coefficient matrix $C$, where

$$C = \begin{bmatrix} \alpha_{0,0} & \alpha_{0,1} & \cdots \\ \alpha_{1,0} & \alpha_{1,1} & \cdots \\ \vdots & \vdots & \\ \alpha_{K-1,0} & \alpha_{K-1,1} & \cdots \end{bmatrix}, \tag{7}$$

must satisfy that every set of $K$ columns of it is a linearly independent set. Once this condition is met, NC is able to achieve the optimal throughput in wireless broadcast scenarios [Yu et al. 2014].

However, it is highly nontrivial to meet this condition, which hinders the implementation of NC. First, to guarantee the linear independence, the sender chooses coefficients randomly from a sufficiently large $\mathbb{F}_q$ [Lucani et al. 2009; Heide et al. 2009] or regularly collects receiver feedback to make online coding decisions [Fragouli et al. 2007]. While a large $\mathbb{F}_q$ incurs heavy computational loads, collecting feedback could be expensive or even impossible in certain circumstances, such as time-division-duplex satellite communications [Lucani et al. 2009]. Second, to enable the decoding, coding coefficients must be attached to each coded packet, which constitute $\lceil K \log_2 q \rceil$ bits of overhead in each transmission. When $q$ is large and $L$ is small, the throughput loss due to the overhead may overwhelm all the other benefits of NC.

These practical shortages of NC can be easily overcome by the proposed supplemented Pascal matrix. By choosing a sufficiently large $p$ and letting $C = H_{p,K}$, we obtain an NC that is both computational friendly (only operations modulo $p$) and feedback-free. Moreover, for the receivers to retrieve the coding coefficients, the sender only needs to attach the index $u$ to the $u$-th packet, rather than attaching the complete coefficients. Furthermore, the additive formula for Pascal matrix may enable efficient progressive coding/decoding algorithms, which could be our future research direction.

**4C.** *Matroid theory.* A matroid $\mathcal{M} = (E, \mathcal{I})$ is a finite collection of elements called the ground set, $E$, paired with its comprehensive set of independent subsets, $\mathcal{I}$. A uniform matroid $U_n^k$ has $|E| = n$ and the property that *any* subset of size $k$ of $E$ is an element of $\mathcal{I}$ and *no* subset of size $k + 1$ is in $\mathcal{I}$. $U_n^k$ is called $q$-representable if there is a $k \times n$ matrix such that every $k$ columns of it are linearly independent over $\mathbb{F}_q$.

**Corollary 4.2** (representability of uniform matroid). *Any uniform matroid $U_n^k$ that satisfies $n \leqslant q + 1$ is $q$-representable by any $n$ columns of $H_{q,k}$.*    □

It is known that any uniform matroid $U_n^k$ that satisfies $n \leqslant q + 1$ is $q$-representable [Oxley 2011; Ball 2015; Reed and Solomon 1960]; one can obtain another construction from Reed–Solomon codes. $H_{q,k}$ is just another, sparse example.

## 5. Conclusion

In this paper, we proposed the supplemented Pascal matrix, whose first $k$ rows are an MDS matrix over $\mathbb{F}_q$ for any prime power $q$ and positive integer $k \leqslant q$. Our construction can be potentially generalized to a framework that enables low-complexity MDS code constructions and encoding/decoding as well. Our matrix can overcome some practical shortages of network coding and, thus, enables high-performance wireless network coded packet broadcast. Our matrix is in agreement with existing results on the representability of uniform matroids, while also providing new insights into this topic. In the future, we intend to study Pascal-based network coding algorithms. We are also interested in applying our results to other fields such as projective geometry and graph theory.

## References

[Ball 2012] S. Ball, "On sets of vectors of a finite vector space in which every subset of basis size is a basis", *J. Eur. Math. Soc.* **14**:3 (2012), 733–748. MR Zbl

[Ball 2015] S. Ball, *Finite geometry and combinatorial applications*, London Mathematical Society Student Texts **82**, Cambridge Univ. Press, 2015. MR Zbl

[Ball and De Beule 2012] S. Ball and J. De Beule, "On sets of vectors of a finite vector space in which every subset of basis size is a basis, II", *Des. Codes Cryptogr.* **65**:1-2 (2012), 5–14. MR Zbl

[El Rouayheb et al. 2010] S. El Rouayheb, A. Sprintson, and C. Georghiades, "On the index coding problem and its relation to network coding and matroid theory", *IEEE Trans. Inform. Theory* **56**:7 (2010), 3187–3195. MR

[Fragouli et al. 2007] C. Fragouli, D. Lun, M. Medard, and P. Pakzad, "On feedback for network coding", pp. 248–252 in *CISS '07*, Annual Conference on Information Sciences and Systems **41**, IEEE, Piscataway, NJ, 2007.

[Heide et al. 2009] J. Heide, M. V. Pedersen, F. H. P. Fitzek, and T. Larsen, "Network coding for mobile devices: systematic binary random rateless codes", pp. 1–6 in *IEEE International Conference on Communications Workshops* (Dresden, 2009), IEEE, Piscataway, NJ, 2009.

[Ho et al. 2006] T. Ho, M. Médard, R. Koetter, D. R. Karger, M. Effros, J. Shi, and B. Leong, "A random linear network coding approach to multicast", *IEEE Trans. Inform. Theory* **52**:10 (2006), 4413–4430. MR Zbl

[Li et al. 2003] S.-Y. R. Li, R. W. Yeung, and N. Cai, "Linear network coding", *IEEE Trans. Inform. Theory* **49**:2 (2003), 371–381. MR Zbl

[Lin and Costello 2004] S. Lin and D. J. Costello, *Error control coding, fundamentals and applications*, 2nd ed., Pearson, Upper Saddle River, NJ, 2004. Errata by E. Agrell available at arXiv 1101.2575. Zbl

[Lucani et al. 2009] D. E. Lucani, M. Medard, and M. Stojanovic, "Random linear network coding for time-division duplexing: field size considerations", pp. 1–6 in *GLOBECOM 2009: IEEE Global Telecommunications Conference* (Honolulu, 2009), edited by M. Ulema, IEEE, Piscataway, NJ, 2009.

[Oxley 2011] J. Oxley, *Matroid theory*, 2nd ed., Oxford Graduate Texts in Mathematics **21**, Oxford Univ. Press, 2011. MR Zbl

[Oxley et al. 1996] J. Oxley, D. Vertigan, and G. Whittle, "On inequivalent representations of matroids over finite fields", *J. Combin. Theory Ser. B* **67**:2 (1996), 325–343. MR Zbl

[Reed and Solomon 1960] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields", *J. Soc. Indust. Appl. Math.* **8**:2 (1960), 300–304. MR Zbl

[Yu et al. 2014] M. Yu, P. Sadeghi, and N. Aboutorab, "On deterministic linear network coded broadcast and its relation to matroid theory", pp. 536–540 in *2014 IEEE Information Theory Workshop* (Hobart, 2014), IEEE, Piscataway, NJ, 2014.

mikwa@umich.edu            Department of Nuclear Engineering and Radiological Sciences
                           and Department of Mathematics, University of Michigan,
                           Ann Arbor, MI, United States

damelin@umich.edu          Mathematical Reviews, The American Mathematical Society,
                           Ann Arbor, MI, United States

jeffjeff@umich.edu         Department of Mathematics, University of Michigan,
                           Ann Arbor, MI, United States

ming.yu@anu.edu.au         College of Engineering and Computer Science,
                           Australian National University, Canberra, Australia

# Hexatonic systems and dual groups in mathematical music theory

Cameron Berry and Thomas M. Fiore

(Communicated by Joseph A. Gallian)

Motivated by the music-theoretical work of Richard Cohn and David Clampitt on late-nineteenth century harmony, we mathematically prove that the *PL*-group of a hexatonic cycle is dual (in the sense of Lewin) to its $T/I$-stabilizer. Our points of departure are Cohn's notions of maximal smoothness and hexatonic cycle, and the symmetry group of the 12-gon; we do *not* make use of the duality between the $T/I$-group and *PLR*-group. We also discuss how some ideas in the present paper could be used in the proof of $T/I$-*PLR* duality by Crans, Fiore, and Satyendra (*Amer. Math. Monthly* **116**:6 (2009), 479–495).

## 1. Introduction: hexatonic cycles and associated dual groups

Why did late nineteenth century composers, such as Franck, Liszt, Mahler, and Wagner, continue to favor consonant triads over other tone collections, while simultaneously moving away from the diatonic scale and classical tonality?

Richard Cohn [1996] proposed an answer, independent of acoustic consonance: major and minor triads are preferred because they can form *maximally smooth cycles*. Consider for instance the following sequence of consonant triads, called a *hexatonic cycle* by Cohn:
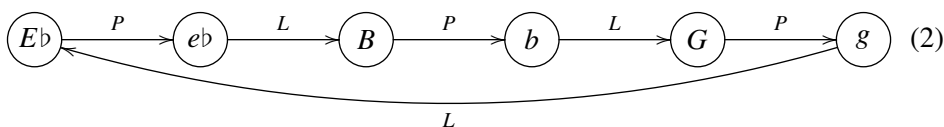
$$E\flat, e\flat, B, b, G, g, E\flat. \tag{1}$$

We have indicated major chords with capital letters and minor chords with lowercase letters. Although the motion from a major chord to its parallel minor, e.g., $E\flat$ to $e\flat$, $B$ to $b$, and $G$ to $g$, is distinctly nondiatonic, this sequence has cogent properties of importance to late-Romantic composers, as axiomatized in Cohn's notion of maximally smooth cycle [1996, page 15]:

- It is a *cycle* in the sense that the first and last chords are the same but all others are different. A *cycle* is required to contain more than three chords.
- All of the chords are in one "set class"; in this case each chord is a consonant triad.
- Every transition is *maximally smooth* in the sense that two notes stay the same while the third moves by the smallest possible interval: a semitone.

Cohn considered movement along this sequence transformationally as an action by a cyclic group of order 6. Additionally, David Clampitt [1998] considered movement along this sequence via $P$ and $L$, and also via certain rotations and reflections. As usual, we denote by $P$ the "parallel" transformation that sends a major or minor chord to its parallel minor or major chord, respectively. We denote by $L$ the "leading tone exchange" transformation, which moves the root of a major chord down a semitone and the fifth of a minor chord up a semitone, so the $L$ sends consonant triads $e\flat$ to $B$, and $b$ to $G$, and $g$ to $E\flat$. The hexatonic cycle (1) is then positioned in the network

$$
\begin{array}{ccccccccccc}
\boxed{E\flat} & \xrightarrow{P} & \boxed{e\flat} & \xrightarrow{L} & \boxed{B} & \xrightarrow{P} & \boxed{b} & \xrightarrow{L} & \boxed{G} & \xrightarrow{P} & \boxed{g}
\end{array} \quad (2)
$$

$$L$$

with alternating $P$ and $L$ transformations between the nodes.

Wagner's Grail motive in *Parsifal* can be interpreted in terms of network (2), as proposed by David Clampitt [1998]. A small part of Clampitt's analysis of the first four chords is pictured in Figure 1. Clampitt includes the final $D\flat$ chord, which lies outside of the hexatonic cycle (1), in his interpretation via a conjugation-modulation applied to a certain subsystem. A third interpretation, in addition to the cyclic one of Cohn [1996, Example 5] and the *PL*-interpretation in Figure 1, was also proposed by Clampitt, this time in terms of the transpositions and inversions $\{T_0, T_4, T_8, I_1, I_5, I_9\}$. Clampitt observes that this group and the *PL*-group form *dual groups in the sense of Lewin* [1987], via their actions on the hexatonic set of chords in (1). The perceptual basis of all three groups is explained in [Clampitt 1998].

The contribution of the present article is to directly prove that the *PL*-group and the group $\{T_0, T_4, T_8, I_1, I_5, I_9\}$ in Clampitt's article are dual groups acting on (1). Our points of departure are the hexatonic cycle (1), the standard action of the dihedral group of order 24 on the 12-gon, and the Orbit-Stabilizer Theorem. We do *not* use the duality of the $T/I$-group and *PLR*-group. Some arguments in Section 3 are similar to arguments of Crans, Fiore, and Satyendra [Crans et al. 2009], but there are important differences; see Remark 3.10.

Just how special are the consonant triads with regard to the maximal smoothness property? According to [Cohn 1996], only six categories of tone collections support
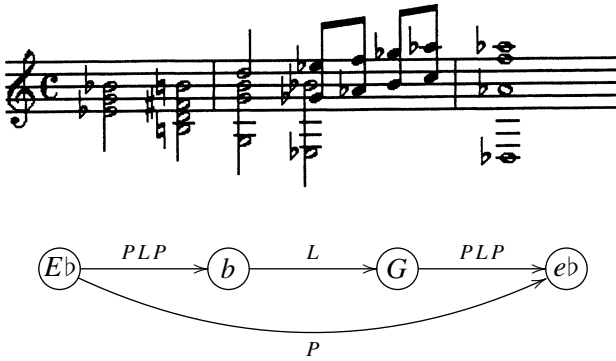
**Figure 1.** Top: Grail motive from Wagner, *Parsifal*, Act 3, measures 1098–1100, reproduced from [Clampitt 1998, Example 1]. Bottom: First four chords of the Grail motive in a hexatonic *PL*-network of Clampitt. Notice that the bottom arrow is the composite of the three top arrows, and goes in the opposite direction of the bottom arrow of diagram (2).

maximally smooth cycles: singletons, consonant triads, pentatonic sets, diatonic sets, complements of consonant triads, and 11-note sets. Clearly the singletons and 11-note sets do not give musically significant cycles. The pentatonic sets and the diatonic sets each support only one long cycle, which exhausts all 12 of their respective exemplars. The consonant triads and their complements, *on the other hand,* support short cycles that do not exhaust all of their transpositions and inversions. The maximally smooth cycles of consonant triads are enumerated as sets as follows:

$$\{E\flat, e\flat, B, b, G, g\}, \tag{3}$$

$$\{E, e, C, c, A\flat, a\flat\}, \tag{4}$$

$$\{F, f, C\sharp, c\sharp, A, a\}, \tag{5}$$

$$\{F\sharp, f\sharp, D, d, B\flat, b\flat\}. \tag{6}$$

These are the four *hexatonic cycles* of Cohn [1996, page 17]. They (and their reverses and complements) are the only short maximally smooth cycles that exist in the Western chromatic scale.

## 2. Mathematical and musical preliminaries: standard dihedral group action on consonant triads and the Orbit-Stabilizer Theorem

We quickly recall the standard preliminaries about consonant triads, transposition, inversion, $P$, $L$, and the Orbit-Stabilizer Theorem. A good introduction to this very

| major triads | minor triads |
|---|---|
| $C = \langle 0, 4, 7 \rangle$ | $\langle 0, 8, 5 \rangle = f$ |
| $C\sharp = D\flat = \langle 1, 5, 8 \rangle$ | $\langle 1, 9, 6 \rangle = f\sharp = g\flat$ |
| $D = \langle 2, 6, 9 \rangle$ | $\langle 2, 10, 7 \rangle = g$ |
| $D\sharp = E\flat = \langle 3, 7, 10 \rangle$ | $\langle 3, 11, 8 \rangle = g\sharp = a\flat$ |
| $E = \langle 4, 8, 11 \rangle$ | $\langle 4, 0, 9 \rangle = a$ |
| $F = \langle 5, 9, 0 \rangle$ | $\langle 5, 1, 10 \rangle = a\sharp = b\flat$ |
| $F\sharp = G\flat = \langle 6, 10, 1 \rangle$ | $\langle 6, 2, 11 \rangle = b$ |
| $G = \langle 7, 11, 2 \rangle$ | $\langle 7, 3, 0 \rangle = c$ |
| $G\sharp = A\flat = \langle 8, 0, 3 \rangle$ | $\langle 8, 4, 1 \rangle = c\sharp = d\flat$ |
| $A = \langle 9, 1, 4 \rangle$ | $\langle 9, 5, 2 \rangle = d$ |
| $A\sharp = B\flat = \langle 10, 2, 5 \rangle$ | $\langle 10, 6, 3 \rangle = d\sharp = e\flat$ |
| $B = \langle 11, 3, 6 \rangle$ | $\langle 11, 7, 4 \rangle = e$ |

**Table 1.** The set of consonant triads, denoted Triads, as displayed on page 483 of [Crans et al. 2009].

well-known background material is [Crans et al. 2009]. Since this background has been treated in many places, we merely rapidly introduce the notation and indicate a few sources.

***Consonant triads.*** We encode pitch classes using the standard $\mathbb{Z}_{12}$ model, where $C = 0$, $C\sharp = D\flat = 1$, and so on, up to $B = 11$. Via this bijection we freely refer to elements of $\mathbb{Z}_{12}$ as *pitch classes*. Major chords are indicated as ordered 3-tuples in $\mathbb{Z}_{12}$ of the form $\langle x, x + 4, x + 7 \rangle$, where $x$ ranges through $\mathbb{Z}_{12}$. Minor chords are indicated as 3-tuples $\langle x + 7, x + 3, x \rangle$ with $x \in \mathbb{Z}_{12}$. We choose these orderings to make simple formulas for $P$ and $L$; this is not a restriction for applications, as the framework was extended in [Fiore et al. 2013a] to allow any orderings. We call the set of 24 major and minor triads Triads, this is the set of *consonant triads*. The letter names are indicated in Table 1.

***Transposition and inversion, and P and L.*** The 12-tone operations *transposition* $T_n : \mathbb{Z}_{12} \to \mathbb{Z}_{12}$ and *inversion* $I_n : \mathbb{Z}_{12} \to \mathbb{Z}_{12}$ are given by

$$T_n(x) = x + n \quad \text{and} \quad I_n(x) = -x + n$$

for $n \in \mathbb{Z}_{12}$. These 24 operations are the symmetries of the regular 12-gon, when we consider 0 through 11 as arranged on the face of a clock. In the music-theory tradition, this group is called the *T/I*-group (the "/" does *not* indicate any kind of quotient). The unique reflection of the 12-gon which interchanges $m$ and $n$ is $I_{m+n}$, as can be verified by direct computation.

Many composers, for instance Schoenberg, Berg, and Webern, utilized these mod 12 transpositions and inversions. These functions and their compositional uses have been thoroughly explored by composers, music theorists, and mathematicians; see for example [Babbitt 1955; Forte 1973; Fripertinger and Lackner 2015; Hook 2007; Hook and Peck 2015; McCartin 1998; Mead 2015; Morris 1987; 1991; 2001; 2015; Rahn 1987]. Indeed, the three recent papers [Fripertinger and Lackner 2015; Mead 2015; Morris 2015] together contain over 100 references.

We consider these bijective functions on $\mathbb{Z}_{12}$ also as bijective functions on Triads via their componentwise evaluation on consonant triads:

$$T_n\langle x_1, x_2, x_3\rangle = \langle T_n x_1, T_n x_2, T_n x_3\rangle \quad \text{and} \quad I_n\langle x_1, x_2, x_3\rangle = \langle I_n x_1, I_n x_2, I_n x_3\rangle. \quad (7)$$

Also on the set Triads of consonant triads (with the indicated ordering), but not on the level of individual pitch classes, we have the bijective functions $P$ and $L$ defined by

$$P\langle x_1, x_2, x_3\rangle = I_{x_1+x_3}\langle x_1, x_2, x_3\rangle \quad \text{and} \quad L\langle x_1, x_2, x_3\rangle = I_{x_2+x_3}\langle x_1, x_2, x_3\rangle. \quad (8)$$

As remarked above, $P$ stands for "parallel" and $L$ stands for "leading tone exchange".

We consider $T_n$, $I_n$, $P$, and $L$ as elements of the symmetric group Sym(Triads).

**Proposition 2.1.** *The bijections $P$ and $L$ commute with $T_n$ and $I_n$ as elements of the symmetric group* Sym(Triads).

*Proof.* This is a straightforward computation using equations (7) and (8). This computation has been discussed in broader contexts in [Fiore et al. 2013b] and [Fiore and Satyendra 2005]. □

***Orbit-Stabilizer Theorem.*** Suppose $S$ is a set with a left group action by a group $G$ (all group actions in this paper are *left* group actions). Recall that the *orbit of an element* $Y \in S$ is

$$\text{orbit of } Y := \{gY \mid g \in G\}.$$

The *stabilizer group of an element* $Y \in S$ is

$$G_Y := \{g \in G \mid gY = Y\}.$$

**Theorem 2.2** (Orbit-Stabilizer Theorem). *Let $G$ be a group with an action on a set $S$. Neither $G$ nor $S$ is assumed to be finite. Then the assignment*

$$G/G_Y \rightarrow \text{orbit of } Y,$$

$$gG_Y \mapsto gY,$$

*is a bijection. In particular, if $G$ is finite, then each orbit is finite, and*

$$|G|/|G_Y| = |\text{orbit of } Y|. \quad (9)$$

*Simple transitivity.* A group action of a group $G$ on a set $S$ is said to be *simply transitive* if for any $Y, Z \in S$ there is a unique $g \in G$ such that $gY = Z$. Informally, we also say the group $G$ is *simply transitive* if the sole action under consideration is simply transitive.

**Proposition 2.3.** (1) *An action of a group $G$ on a set $S$ is simply transitive if and only if it is transitive and every stabilizer $G_Y$ is trivial.*

(2) *Suppose $G$ is a finite group that acts on a set $S$. Then $G$ is simply transitive if and only if any two of the following three hold:*

   (a) *$G$ is transitive.*
   (b) *Every stabilizer $G_Y$ is trivial.*
   (c) *$G$ and $S$ have the same cardinality.*

   *In this case, the third condition also holds.*

   *Another way to read this "if and only if" statement is: assuming $G$ is finite and any one of the conditions holds, $G$ is simply transitive if and only if another one of the conditions holds.*

(3) *Suppose a (not necessarily finite) group $H_1$ acts simply transitively on a set $S$, and a subgroup $H_2$ of $H_1$ acts transitively on $S$ via its subaction. Then $H_1 = H_2$.*

*Proof.* (1) If the action is simply transitive, then it acts transitively and for each $Y \in S$, there is only one $g \in G$ with $gY = Y$, and hence each $G_Y$ is trivial.

Suppose $G$ acts transitively and for every $Y \in S$, the group $G_Y$ is trivial. Suppose $Y, Z \in S$ and $g_1, g_2 \in G$ satisfy $g_1Y = Z$ and $g_2Y = Z$. Then $Y = g_2^{-1}Z$ and $g_2^{-1}g_1Y = Y$, so $g_2^{-1}g_1 \in G_Y = \{e\}$, and finally $g_1 = g_2$.

(2) We first prove that any two of the conditions implies the third and implies simple transitivity.

(a)(b) $\Rightarrow$ (c): $G$ is simply transitive by (1), and equation (9) says $|G|/1 = |S|$, so $|G| = |S|$ and (c) holds.

(b)(c) $\Rightarrow$ (a): Equation (9) says $|S| = |G|/1 = |$orbit of $Y|$, so $S =$ orbit of $Y$, and $G$ is transitive and (a) holds, so $G$ is simply transitive by (1).

(a)(c) $\Rightarrow$ (b): Equation (9) says $|G|/|G_Y| = |G|$, so $|G_Y| = 1$ and (b) holds, and $G$ is simply transitive by (1).

Now that we have shown any two of the conditions implies the third and simple transitivity, we want to see that simply transitivity implies all three conditions. From (1), simple transitivity implies (a) and (b), and we have already seen (a) and (b) imply (c).

(3) Suppose $H_1$ properly contains $H_2$, and $h_1 \in H_1 \backslash H_2$. Fix a $Y \in S$ and define $Z := h_1Y$. Then by the transitivity of $H_2$, there is an $h_2 \in H_2$ such that $Z = h_2Y$. But by the simple transitivity of $H_1$, we must have $h_1 = h_2$, a contradiction.  □

## 3. Main theorem: Hexatonic Duality

We next review the notion of dual groups, and then turn to the main result, Theorem 3.9 on hexatonic duality. Recall that subgroups $G$ and $H$ of $\mathrm{Sym}(S)$ are *dual in the sense of Lewin* [1987, page 253] if each acts simply transitively on $S$ and each is the centralizer of the other.[1] Recall the *centralizer of $G$ in* $\mathrm{Sym}(S)$ is

$$C(G) = \{\sigma \in \mathrm{Sym}(S) \mid \sigma g = g\sigma \text{ for all } g \in G\}.$$

Before turning to the main result, we prove two simultaneous redundancies in the notion of *dual groups*: instead of requiring the two groups to centralize each other, it is sufficient to merely require that they commute, and instead of requiring $H$ to act simply transitively, it is sufficient to merely require $H$ acts transitively.

**Proposition 3.1.** *Let $S$ be a (not necessarily finite) set. Suppose $G \le \mathrm{Sym}(S)$ acts simply transitively on $S$ and $H \le \mathrm{Sym}(S)$ acts transitively on $S$. Suppose $G$ and $H$ commute in the sense that $gh = hg$ for all $g \in G$ and $h \in H$. Then $G$ and $H$ are dual groups. In particular, $H$ also acts simply transitively and $G$ and $H$ centralize one another.*

*Proof.* We would like to first conclude from the simple transitivity of $G$, the transitivity of $H$, and the commutativity of $G$ and $H$, that the centralizer $C(G)$ acts simply transitively on $S$.

We claim $C(G)$ acts simply transitively on $S$. It acts transitively, as $C(G) \supseteq H$ and $H$ acts transitively. So, it suffices by Proposition 2.3(1) to prove that, for each $s \in S$, the only element of $C(G)$ that fixes $s$ is the identity. Let $\sigma$ be an element of $C(G)$ that fixes $s$, and $g$ any element of $G$. Then,

$$\sigma s = s \implies g(\sigma s) = g(s) \implies (g\sigma)s = (gs) \implies (\sigma g)s = (gs) \implies \sigma(gs) = (gs).$$

---

[1]Lewin [1987, page 253] gave a more general situation that gives rise to examples of dual groups in the sense defined above, though he did not formally make this definition. He starts with a group $G$, there called *STRANS*, assumed to act simply transitively on a set $S$, and then makes three claims without proof: (1) the centralizer $C(G)$ in $\mathrm{Sym}(S)$ acts simply transitively on $S$ (the centralizer $C(G)$ is called *STRANS′* there); (2) the double centralizer $C(C(G))$ is contained in $G$, so actually $C(C(G)) = G$; and (3) the two generalized interval systems with transposition groups $G$ and $C(G)$ respectively have interval-preserving transformation groups, precisely $C(G)$ and $G$ respectively. See Proposition 3.2 for a proof of statements (1) and (2). Statement (3) is a consequence of the first two statements in combination with *COMM-SIMP* duality, which was stated on page 101 of [Lewin 1995] and partially proved in [Lewin 1987, Theorem 3.4.10]. For a review of *COMM-SIMP* duality and another proof, see [Fiore and Satyendra 2005, Section 2 and Appendix]. For the equivalence of generalized interval systems and simply transitive group actions, see pages 157–159 of Lewin's monograph. The equivalence on the level of categories was proved by Fiore, Noll and Satyendra [Fiore et al. 2013b, page 10]. The undergraduate research project of Sternberg [2006] worked out some of the details of Lewin's simply transitive group action associated to a generalized interval system and investigated the Fugue in F from Hindemith's *Ludus Tonalis*.

So, not only does $\sigma$ fix $s$, but $\sigma$ also fixes $(gs)$ for every $g \in G$. That is to say, $\sigma = \mathrm{Id}_S$, and $C(G)$ acts simply transitively on $S$.

Now we have the transitive subgroup $H$ contained in the simply transitive group $C(G)$ by the assumed commutativity, so by Proposition 2.3(3), $H = C(G)$, and $H$ also acts simply transitively.

To obtain $C(H) = G$, we use the newly achieved simple transitivity of $H$ and repeat the argument with the roles of $G$ and $H$ reversed. $\qquad\square$

We may now use a result of Dixon and Mortimer to prove a statement of Lewin [1987, page 253], as suggested by Julian Hook, Robert Peck, and Thomas Noll. Parts (1) and (2) of the following proposition were stated by Lewin.

**Proposition 3.2.** *Let $S$ be a* (*not necessarily finite*) *set. Suppose $G \leq \mathrm{Sym}(S)$ acts simply transitively on $S$. Then*:

(1) *The centralizer $C(G)$ in $\mathrm{Sym}(S)$ acts simply transitively on $S$.*

(2) *The centralizer of the centralizer $C(C(G))$ is equal to $G$.*

(3) *Define $H := C(G)$. Then $G$ and $H$ are dual groups.*

*Proof.* (1) This follows immediately from [Dixon and Mortimer 1996, Theorem 4.2A(i) and (ii), page 109]. There *semiregular* means point stabilizers are trivial and *regular* means simply transitive.

(2) Since $C(G)$ is simply transitive, we can apply Dixon and Mortimer's result to $C(G)$ to get that the double centralizer $C(C(G))$ simply transitive. But $C(C(G))$ contains the simply transitive group $G$, so $C(C(G)) = G$ by Proposition 2.3(3).

(3) This follows directly from the preceding two by definition. $\qquad\square$

We now turn to the discussion of our main result.

Let Hex be the set of chords in the hexatonic cycle (1) and $\underline{\mathrm{Hex}}$ the set of underlying pitch classes of its chords; that is,

$$\mathrm{Hex} := \{E\flat, e\flat, B, b, G, g\}, \quad \underline{\mathrm{Hex}} := \{2, 3, 6, 7, 10, 11\}.$$

Our goal is to prove that the restriction of the *PL*-group to Hex and the restriction of $\{T_0, T_4, T_8, I_1, I_5, I_9\}$ to Hex are dual groups, and that each is dihedral of order 6. The strategy is to separately prove the unrestricted groups act simply transitively and are dihedral, and then finally to show that the restricted groups centralize each other. We begin with a characterization of the consonant triads contained in $\underline{\mathrm{Hex}}$.

**Lemma 3.3.** *The only consonant triads of Table 1 contained in $\underline{\mathrm{Hex}}$ as subsets are the elements of* Hex.

*Proof.* We first identify the available perfect fifths in $\underline{\mathrm{Hex}}$ (pairs with difference 7), and then check if the corresponding major/minor thirds are in $\underline{\mathrm{Hex}}$.
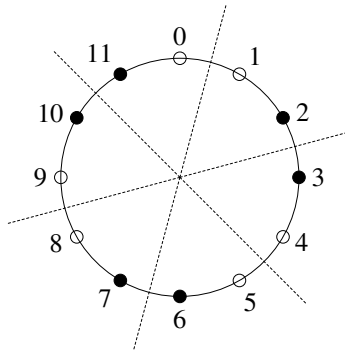
**Figure 2.** The solid circles represent the subset $\underline{\text{Hex}}$ of $\mathbb{Z}_{12}$. The symmetry of the subset makes apparent that the only rotations which preserve $\underline{\text{Hex}}$ are $T_0$, $T_4$, and $T_8$. The geometric locations of the solid circles also imply that the reflections across the dashed lines are the only reflections which preserve $\underline{\text{Hex}}$.

The only pairs of the form $\langle x, x+7 \rangle$ are $\langle 3, 10 \rangle$, $\langle 7, 2 \rangle$, and $\langle 11, 6 \rangle$, and we see that $x + 4$ is contained in $\underline{\text{Hex}}$ in each case; that is, 7, 11, and 3 are in $\underline{\text{Hex}}$. Thus we have the three major chords $E\flat$, $G$, and $B$, and no others.

The only pairs of the form $\langle x+7, x \rangle = \langle y, y+5 \rangle$ are $\langle 2, 7 \rangle$, $\langle 6, 11 \rangle$, and $\langle 10, 3 \rangle$, and we see that $x + 3 = y + 8$ is contained in $\underline{\text{Hex}}$ in each case; that is, 10, 2, and 6 are in $\underline{\text{Hex}}$. Thus we have the three minor chords $g$, $b$, and $e\flat$, and no others.  □

**Proposition 3.4.** (1) *The only elements of the $T/I$-group that preserve* Hex *as a set are* $\{T_0, T_4, T_8, I_1, I_5, I_9\}$, *so they form a group, which we will denote by H.*

(2) $H := \{T_0, T_4, T_8, I_1, I_5, I_9\}$ *is dihedral of order* 6.

*Proof.* (1) If an element of the $T/I$-group preserves Hex as a set, then it must also preserve the collection $\underline{\text{Hex}}$ of underlying pitch classes as a set. Geometric inspection of the plot of $\underline{\text{Hex}}$ in Figure 2 reveals that the only rotations that preserve $\underline{\text{Hex}}$ are $T_0$, $T_4$, and $T_8$.

Again looking at Figure 2, we see that the three reflections which interchange $2 \leftrightarrow 3$ or $6 \leftrightarrow 7$ or $10 \leftrightarrow 11$ preserve $\underline{\text{Hex}}$. By a comment on page 256, these are

$$I_{2+3} = I_5, \quad I_{6+7} = I_1, \quad I_{10+11} = I_9.$$

No other reflections preserve $\underline{\text{Hex}}$, as we can see geometrically from its limited reflection symmetry.

Since $H := \{T_0, T_4, T_8, I_1, I_5, I_9\}$ is a setwise stabilizer of $\underline{\text{Hex}}$, it is a group.

From Lemma 3.3 we see that $\{T_0, T_4, T_8, I_1, I_5, I_9\}$ must also stabilize the chord collection Hex as a set. No other transpositions or inversions stabilize Hex by the argument at the outset of this proof.

(2) The only noncommutative group of order 6 is the symmetric group on three elements, denoted Sym(3), which is isomorphic to the dihedral group of order 6. The group under consideration is noncommutative, because $T_4 I_1(x) = -x + 5$ while $I_1 T_4(x) = -x - 3$. □

**Proposition 3.5.** *The setwise stabilizer* $H$ *acts simply transitively on* Hex.

*Proof.* The $H$-orbit of $E\flat$ is all of Hex, as the following diagram shows.
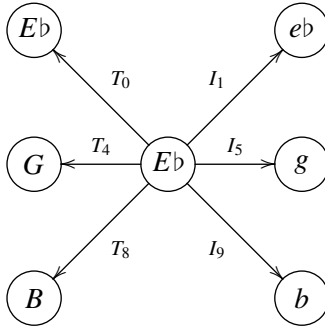


We have $|H| = 6 = |$orbit of $Y|$ so the Orbit-Stabilizer Theorem

$$|H|/|H_Y| = |\text{orbit of } Y|$$

implies $|H_Y| = 1$. See Proposition 2.3(2). □

Next we can investigate the subgroup of Sym(Triads) generated by $P$ and $L$, which is called the *PL-group*.

**Proposition 3.6.** *The subgroup* $\langle P, L \rangle$ *of* Sym(Triads) *is dihedral of order* 6.

*Proof.* We first observe that $P$ and $L$ are involutions; that is, $P^2 = \text{Id}_{\text{Triads}}$ and $L^2 = \text{Id}_{\text{Triads}}$. A musical justification comes from the definitions of "parallel" and "leading tone exchange". Direct computations of $P^2$ and $L^2$ using the formulas in (8) provide a mathematical justification.

Since $P$ and $L$ are involutions, every nontrivial element of $\langle P, L \rangle$ can be expressed as an alternating word in the letters $P$ and $L$. The six functions $\text{Id}_{\text{Triads}}$, $P$, $LP$, $PLP$, $LPLP$, and $PLPLP$ are all distinct by evaluating at $E\flat$ using the following diagram from the Introduction.
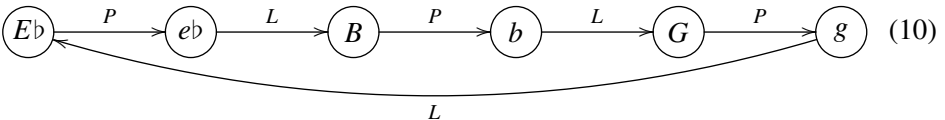


From diagram (10) we also see that $(LP)^3(E\flat) = E\flat$, and for any $Y \in \{E\flat, B, G\}$, $(LP)^3(Y) = Y$. Similarly, by reading the diagram backwards (recall $P$ and $L$ are involutions), we see $(LP)^3(Y) = Y$ for any minor triad $Y \in \{e\flat, b, g\}$. We have similar *PL*-diagrams and considerations for the cycles in (4), (5), and (6), and

therefore $(LP)^3 = \mathrm{Id}_{\mathrm{Triads}}$ on the entire set Triads of consonant triads. Another way to see that $(LP)^3 = \mathrm{Id}_{\mathrm{Triads}}$ is to combine the observation $(LP)^3(E\flat) = E\flat$ from diagram (10) with Proposition 2.1 and the fact that Triads is the $T/I$-orbit of $E\flat$.

We next show via a word-theoretic argument that $\langle P, L \rangle$ consists only of the six functions $\mathrm{Id}_{\mathrm{Triads}}$, $P$, $LP$, $PLP$, $LPLP$, and $PLPLP$ discussed above. From $(LP)^3 = \mathrm{Id}_{\mathrm{Triads}}$, we express $PL$ in terms of $LP$. Namely,

$$(LP)^3 = \mathrm{Id}_{\mathrm{Triads}} \implies (LP)^3(PL) = (PL) \implies (LP)^2 = PL.$$

Consider any alternating word in $P$ and $L$. If the rightmost letter is $P$, then we can use $(LP)^3 = \mathrm{Id}_{\mathrm{Triads}}$ to achieve an equality with one of the six functions we already have. If the rightmost letter is $L$, then we replace each $PL$ by $(LP)^2$ and use $L^2 = \mathrm{Id}_{\mathrm{Triads}}$ if $LL$ results on the far left. Then we have an equal function with rightmost letter $P$, which we can then reduce to one of the six above using $(LP)^3 = \mathrm{Id}_{\mathrm{Triads}}$, as we did in the first case of rightmost letter $P$. Thus $\langle P, L \rangle = \{\mathrm{Id}_{\mathrm{Triads}}, P, LP, PLP, LPLP, PLPLP\}$.

This group is noncommutative, as $PL \neq LP$; hence it is isomorphic to $\mathrm{Sym}(3)$, the only noncommutative group of order 6. But $\mathrm{Sym}(3)$ is dihedral of order 6.

Instead of the previous paragraph, we can show $\langle P, L \rangle$ is dihedral of order 6 using a presentation. Let $t := L$ and $s := LP$; then $s^3 = e$, $t^2 = e$, and $tst = s^{-1}$. The dihedral group of order 6 is the largest group with elements $s$ and $t$ such that $s^3 = e$, $t^2 = e$, and $tst = s^{-1}$. But we observed from diagram (10) that $\langle P, L \rangle$ has at least six distinct elements. Hence, $\langle P, L \rangle$ is dihedral of order 6. □

**Proposition 3.7.** *The PL-group $\langle P, L \rangle$ acts simply transitively on* Hex.

*Proof.* From diagram (10) we see that $\langle P, L \rangle$ acts transitively on Hex. Since $\langle P, L \rangle$ and Hex have the same cardinality, the Orbit-Stabilizer Theorem implies that every stabilizer must be trivial. See Proposition 2.3(2). □

**Lemma 3.8.** *Let $S$ be a set and suppose $G \leq \mathrm{Sym}(S)$. Suppose $G$ acts simply transitively on an orbit $\bar{S}$, and $\bar{G}$ is the restriction of $G$ to the orbit $\bar{S}$. Then the restriction homomorphism $G \to \bar{G}$ is an isomorphism, and $\bar{G}$ also acts simply transitively.*

*Proof.* Suppose $g \in G$ has restriction $\bar{g}$ with $\bar{g}\bar{s} = \bar{s}$ for all $\bar{s} \in \bar{S}$. Then $g$ also has $g\bar{s} = \bar{s}$ for all $\bar{s} \in \bar{S}$, so $g = \mathrm{Id}_S$ by simple transitivity, and the kernel of the surjective homomorphism $G \to \bar{G}$ is trivial. The transitivity of $\bar{G}$ is clear: for any $\bar{s}, \bar{t} \in \bar{S}$ there exists $g \in G$ such that $g\bar{s} = \bar{t}$, so also $\bar{g}\bar{s} = \bar{t}$ with $\bar{g} \in \bar{G}$. The uniqueness of $\bar{g} \in \bar{G}$ is also clear: if $\bar{h} \in \bar{G}$ also satisfies $\bar{h}\bar{s} = \bar{t}$, then so do $g$ and $h$, so $g = h$ by the simple transitivity of $G$ acting on $\bar{S}$, so $\bar{g} = \bar{h}$. □

**Theorem 3.9** (Hexatonic Duality). *The restrictions of the PL-group and the group $H = \{T_0, T_4, T_8, I_1, I_5, I_9\}$ to* Hex *are dual groups in* $\mathrm{Sym}(\mathrm{Hex})$, *and both are dihedral of order* 6.

*Proof.* Let $\overline{G}$ be the restriction of the *PL*-group to Hex, and let $\overline{H}$ be the restriction of $H = \{T_0, T_4, T_8, I_1, I_5, I_9\}$ to Hex.

We already know that $\overline{G}$ and $\overline{H}$ are dihedral of order 6 by Propositions 3.4 and 3.6 and Lemma 3.8.

We also already know that $\overline{G}$ and $\overline{H}$ each act simply transitively on Hex by Propositions 3.5 and 3.6 and Lemma 3.8. We even already know that the groups $\overline{G}$ and $\overline{H}$ commute by Proposition 2.1. Finally, Proposition 3.1 guarantees that $\overline{G}$ and $\overline{H}$ centralize one another.                                                          □

**Remark 3.10** (Comparison with the proof strategy of Crans, Fiore, and Satyendra). There are several differences between the proof strategy of hexatonic duality in the present Theorem 3.9 and the proof strategy of *T/I-PLR* duality in Theorem 6.1 of [Crans et al. 2009]. In the present paper, we first proved that the concerned groups act simply transitively, and determined their structure, and only then showed that the groups exactly centralize each other. In [Crans et al. 2009], on the other hand, the determination of the size of the *PLR*-group was postponed until after the centralizer $C(T/I)$ was seen to act simply, i.e., that each stabilizer $C(T/I)_Y$ is trivial. Then, from these trivial stabilizers, the Orbit-Stabilizer Theorem, the earlier observation that $24 \leq |PLR\text{-group}|$, and the consequence

$$24 \leq |PLR\text{-group}| \leq |C(T/I)| \leq |\text{orbit of } Y| \leq 24$$

on page 492, the authors of [Crans et al. 2009] simultaneously conclude that the *PLR*-group has 24 elements and is the centralizer of $T/I$.

A slight simplification of the aforementioned inequality would be an argument like the one in the present paper: observe that the *PLR*-group acts transitively on the 24 consonant triads because of the Cohn *LR*-sequence, recalled on page 487 of [Crans et al. 2009]; then $C(T/I)$ must act transitively as it contains the *PLR*-group, and then the Orbit-Stabilizer Theorem and the trivial stabilizers imply that $|C(T/I)|$ must be 24, so the *PLR*-group also has 24 elements. Also, instead of postponing the proof that the *PLR*-group has exactly 24 elements from Theorem 5.1 of [Crans et al. 2009] until the aforementioned inequality in Theorem 6.1, one could do a word-theoretic argument in Theorem 5.1 to see that the *PLR*-group has exactly 24 elements, similar to the present argument in Proposition 3.6.

**Remark 3.11.** For an explicit computation of the four hexatonic cycles as orbits of the *PL*-group, see [Oshita 2009], which was also an undergraduate research project with the second author of the present article. That preprint includes a sketch that $\langle P, L \rangle \cong \text{Sym}(3)$.

**Remark 3.12** (Alternative derivation using the Sub Dual Group Theorem). Hexatonic Duality, Theorem 3.9, can also be proved using the Sub Dual Group Theorem of Fiore and Noll [2011, Theorem 3.1], *if one assumes already* the duality of the

| $k$ | $k\mathrm{Hex}$ | $kHk^{-1} =$ dual group to $PL$-group on $k\mathrm{Hex}$ |
|---|---|---|
| $\mathrm{Id}_{\mathrm{Triads}}$ | $\{E\flat, e\flat, B, b, G, g\}$ | $H = \{T_0, T_4, T_8, I_1, I_5, I_9\}$ |
| $T_1$ | $\{E, e, C, c, A\flat, a\flat\}$ | $\{T_0, T_4, T_8, I_3, I_7, I_{11}\}$ |
| $T_2$ | $\{F, f, C\sharp, c\sharp, A, a\}$ | $\{T_0, T_4, T_8, I_5, I_9, I_1\}$ |
| $T_3$ | $\{F\sharp, f\sharp, D, d, B\flat, b\flat\}$ | $\{T_0, T_4, T_8, I_7, I_{11}, I_3\}$ |

**Table 2.** The four hexatonic cycles as $PL$-orbits and the respective dual groups determined as conjugations of $H$ via the Sub Dual Group Theorem of Fiore and Noll.

$T/I$-group and $PLR$-group (maximal smoothness is not discussed in that paper). In Section 3.1 of the same paper, they apply the Sub Dual Group Theorem to the construction of dual groups on the hexatonic cycles. The method is to select $G_0$ to be the $PL$-group, select $s_0 = E\flat$, and compute $S_0 := G_0 s_0 = \mathrm{Hex}$, and then the dual group will consist of the restriction of those elements of the $T/I$-group that map $E\flat$ into $S_0$.

Notice that in the present paper, on the other hand, we *first* determined which transpositions and inversions preserve Hex in Proposition 3.4, and then proved duality, whereas the application of the Sub Dual Group Theorem of Fiore and Noll starts with the $PL$-group and determines from it the dual group as (the restrictions of) those elements of the $T/I$-group that map $E\flat$ into $S_0$. Notice also, in the present paper we determined that the $PL$-group and its dual $H$ are dihedral of order 6, but that the Sub Dual Group Theorem of Fiore and Noll does not specify which group structure is present. In any case, Clampitt [1998] explicitly wrote down all 6 elements of each group in permutation cycle notation.

The present paper is complementary to the work [Fiore and Noll 2011] in that we work very closely with the specific details of the groups and sets involved to determine one pair of dual groups in an illustrative way, rather than appealing to a computationally and conceptually convenient theorem. Fiore and Noll, however, also use their Corollary 3.3 to compute the other hexatonic duals via conjugation, as summarized in Table 2.

The application of the Sub Dual Group Theorem to construct dual groups on octatonic systems is also treated in [Fiore and Noll 2011], and utilized in [Fiore et al. 2013b].

**Remark 3.13** (Other sources on group actions). Music-theoretical group actions on chords have been considered by many, many authors over the past century. In addition to the selected references of Babbitt, Forte, and Morris above, we also mention the expansive and influential work of Mazzola [1985; 1990; 2002] and numerous collaborators. Moreover, issue 42:2 of the *Journal of Music Theory* from

1998 is illuminating obligatory reading on groups in neo-Riemannian theory. That issue contains Clampitt's article [1998], which is the inspiration for the present paper. Clough's article [1998] in that issue illustrates the dihedral group of order 6 and its recombinations with certain centralizer elements in terms of two concentric equilateral triangles (but it does not treat hexatonic systems and duality). The dihedral group of order 6 is a warm-up for his treatment of recombinations of the *Schritt-Wechsel* group with the *T/I*-group, which are both dihedral of order 24. Peck's article [2010] studies centralizers where the requirement of simple transitivity is relaxed in various ways, covering many examples from music theory. Peck determines the structure of centralizers in several cases.

**Remark 3.14** (Discussion of local diatonic containment of hexatonic cycles). No hexatonic cycle is contained entirely in a single diatonic set, as one can see from any of the cycles (3)–(6). However, one can consider a sequence of diatonic sets that changes along with the hexatonic cycle and contains each respective triad, as in [Douthett 2008, Table 4.7]. After transposing and reversing Douthett's table, we see a sequence of diatonic sets such that each diatonic set contains the respective triad of (3).

| triad | $E\flat$ | $e\flat$ | $B$ | $b$ | $G$ | $g$ |
|---|---|---|---|---|---|---|
| in scale | $E\flat$-major | $D\flat$-major | $B$-major | $A$-major | $G$-major | $F$-major |

This sequence of diatonic sets (indicated via major scales) descends by a whole step each time, so is as evenly distributed as possible.

Other diatonic set sequences also contain the hexatonic cycle, though unfortunately there is no maximally smooth cycle of diatonic sets that does the job (recall that the diatonic sets can only form a cycle of length 12). But it is possible to have a maximally smooth sequence of diatonic sets that covers four hexatonic triads. We list all possible diatonic sets containing the respective hexatonic chords.[2]

| triad | $E\flat$ | $e\flat$ | $B$ | $b$ | $G$ | $g$ |
|---|---|---|---|---|---|---|
| in major scales | $E\flat$ $B\flat$ $A\flat$ | $G\flat$ $D\flat$ $B$ | $B$ $F\sharp$ $E$ | $D$ $A$ $G$ | $G$ $D$ $C$ | $B\flat$ $F$ $E\flat$ |

---

[2]Recall that major chords only occur with roots on major scale degrees 1, 4, and 5, so we determine in the table the scales containing a given major triad by considering the root, a perfect fourth below the root, and a perfect fifth below the root. Minor chords can only occur with roots on major scale degrees 2, 3, and 6, so we determine in the table the scales containing a given minor triad by considering a major sixth below the root, a whole step below the root, and a major third below the root. This inconsistent major/minor ordering allows us to see (at vertical dividing lines) all three maximally smooth transitions from diatonic sets containing a given a minor triad to a diatonic set containing its subsequent major in a hexatonic cycle.

Vertical dividing lines indicate maximally smooth transitions between consecutive diatonic sets. As indicated by these dividing lines, the transition from a minor triad to its subsequent major in a hexatonic cycle via $L$ is contained in three maximally smooth transitions of diatonic sets. On the other hand, the transition from a major triad to its subsequent minor in a hexatonic cycle via $P$ is contained in only one maximally smooth transition of diatonic sets, as indicated by the bold letters. Altogether, we can trace three maximally smooth chains of four major scales that contain part of the hexatonic cycle (3):

$$B - E - A - D,$$
$$G - C - F - B\flat,$$
$$E\flat - A\flat - D\flat - F\sharp.$$

Local containment of hexatonic cycles in diatonic chains has ramifications for music analysis. Jason Yust [2013; 2015] proposed to include diatonic contexts into analyses involving $PL$-cycles or $PR$-cycles, and he provides analytical tools to do so.

## 4. Conclusion

We began this article with Cohn's proposal that the maximal smoothness of consonant triads is a key factor for their privileged status in late-nineteenth century music. Indeed, consonant triads and their complements are the only tone collections that accommodate short maximally smooth cycles. The four maximally smooth cycles of consonant triads, the so-called hexatonic cycles of Cohn, can be described transformationally as alternating applications of the neo-Riemannian "parallel" and "leading tone exchange" transformations. Cohn interpreted Wagner's Grail motive in terms of a cyclic group action on the hexatonic cycle Hex, whereas Clampitt used the $PL$-group and the transposition-inversion subgroup we called $H$ in Proposition 3.4. In the present article, we proved the Lewinian duality between these latter two groups, which was discussed by Clampitt [1998].

For perspective, we mention that simply transitive group actions correspond to the *generalized interval systems* of Lewin; see the very influential original source [Lewin 1987], or see [Fiore et al. 2013b, Section 2] for an explanation of some aspects. Dual groups correspond to dual generalized interval systems: the transpositions of one system are the interval-preserving bijections of the other. Clampitt [1998] explained the coherent perceptual basis of the three generalized interval systems associated to the three group actions on Hex by Cohn's cyclic group, the $PL$-group, and the $H$ group. He employed the coherence of generalized interval systems to incorporate the final $D\flat$ of the Grail motive into his interpretation via a conjugation-modulation of a subsystem.

## Acknowledgements

## References

[Babbitt 1955]  M. Babbitt, "Some aspects of twelve-tone composition", *Score IMA Mag.* **12** (June 1955), 53–61.

[Clampitt 1998]  D. Clampitt, "Alternative interpretations of some measures from *Parsifal*", *J. Music Theory* **42**:2 (1998), 321–334.

[Clough 1998]  J. Clough, "A rudimentary geometric model for contextual transposition and inversion", *J. Music Theory* **42**:2 (1998), 297–306.

[Cohn 1996]  R. Cohn, "Maximally smooth cycles, hexatonic systems, and the analysis of late-Romantic triadic progressions", *Music Analysis* **15**:1 (1996), 9–40.

[Crans et al. 2009]  A. S. Crans, T. M. Fiore, and R. Satyendra, "Musical actions of dihedral groups", *Amer. Math. Monthly* **116**:6 (2009), 479–495. MR Zbl

[Dixon and Mortimer 1996]  J. D. Dixon and B. Mortimer, *Permutation groups*, Graduate Texts in Mathematics **163**, Springer, New York, 1996. MR Zbl

[Douthett 2008]  J. Douthett, "Filtered point-symmetry and dynamical voice leading", pp. 72–106 in *Music theory and mathematics: chords, collections, and transformations*, edited by J. Douthett et al., Eastman Studies in Music **50**, Univ. Rochester Press, 2008.

[Fiore and Noll 2011]  T. M. Fiore and T. Noll, "Commuting groups and the topos of triads", pp. 69–83 in *Mathematics and computation in music* (Paris, June, 2011), edited by C. Agon et al., Lecture Notes in Comput. Sci. **6726**, Springer, Heidelberg, 2011. MR Zbl

[Fiore and Satyendra 2005]  T. M. Fiore and R. Satyendra, "Generalized contextual groups", *Music Theory Online* **11**:3 (2005), art. id. 11.3.1.

[Fiore et al. 2013a]  T. M. Fiore, T. Noll, and R. Satyendra, "Incorporating voice permutations into the theory of neo-Riemannian groups and Lewinian duality", pp. 100–114 in *Mathematics and computation in music* (Montreal, June, 2013), edited by J. Yust et al., Lecture Notes in Comput. Sci. **7937**, Springer, Heidelberg, 2013. MR Zbl

[Fiore et al. 2013b]  T. M. Fiore, T. Noll, and R. Satyendra, "Morphisms of generalized interval systems and *PR*-groups", *J. Math. Music* **7**:1 (2013), 3–27. MR Zbl

[Forte 1973]  A. Forte, *The structure of atonal music*, Yale Univ. Press, New Haven, CT, 1973.

[Fripertinger and Lackner 2015]  H. Fripertinger and P. Lackner, "Tone rows and tropes", *J. Math. Music* **9**:2 (2015), 111–172. MR Zbl

[Hook 2007]  J. Hook, "Why are there twenty-nine tetrachords? A tutorial on combinatorics and enumeration in music theory", *Music Theory Online* **13**:4 (2007), art. id. 13.4.1.

[Hook and Peck 2015]  J. Hook and R. Peck, "Introduction to the special issue on tone rows and tropes", *J. Math. Music* **9**:2 (2015), 109–110.

[Lewin 1987]  D. Lewin, *Generalized musical intervals and transformations*, Yale Univ. Press, New Haven, CT, 1987.

[Lewin 1995]  D. Lewin, "Generalized interval systems for Babbitt's lists, and for Schoenberg's String Trio", *Music Theory Spectrum* **17**:1 (1995), 81–118.

[Mazzola 1985]  G. Mazzola, *Gruppen und Kategorien in der Musik: Entwurf einer mathematischen Musiktheorie*, Research and Exposition in Mathematics **10**, Heldermann Verlag, Berlin, 1985. MR Zbl

[Mazzola 1990]  G. Mazzola, *Geometrie der Töne: Elemente der mathematischen Musiktheorie*, Birkhäuser, Basel, 1990. Zbl

[Mazzola 2002]  G. Mazzola, *The topos of music*, Birkhäuser, Basel, 2002. MR Zbl

[McCartin 1998]  B. J. McCartin, "Prelude to musical geometry", *College Math. J.* **29**:5 (1998), 354–370. MR Zbl

[Mead 2015]  A. Mead, "Remarks on "Tone rows and tropes" by Harald Fripertinger and Peter Lackner", *J. Math. Music* **9**:2 (2015), 173–178. MR

[Morris 1987]  R. Morris, *Composition with pitch-classes: a theory of compositional design*, Yale Univ. Press, New Haven, CT, 1987.

[Morris 1991]  R. Morris, *Class notes for atonal music theory*, Frog Peak Music, Hanover, NH, 1991.

[Morris 2001]  R. Morris, *Class notes for advanced atonal music theory*, Frog Peak Music, Lebanon, NH, 2001.

[Morris 2015]  R. Morris, "Review of "Tone rows and tropes" by Harald Fripertinger and Peter Lackner", *J. Math. Music* **9**:2 (2015), 179–195. MR

[Oshita 2009]  K. Oshita, "The hexatonic systems under neo-Riemannian theory: an exploration of the mathematical analysis of music", REU project, Department of Mathematics, University of Chicago, 2009, available at http://tinyurl.com/oshita.

[Peck 2010]  R. Peck, "Generalized commuting groups", *J. Music Theory* **54**:2 (2010), 143–177.

[Rahn 1987]  J. Rahn, *Basic atonal theory*, Schirmer, New York, 1987.

[Sternberg 2006]  J. Sternberg, "Conceptualizing music through mathematics and the generalized interval system", REU project, Department of Mathematics, University of Chicago, 2006, available at http://tinyurl.com/sternbergreu.

[Yust 2013]  J. Yust, "Tonal prisms: iterated quantization in chromatic tonality and Ravel's 'Ondine'", *J. Math. Music* **7**:2 (2013), 145–165. MR Zbl

[Yust 2015]  J. Yust, "Distorted continuity: chromatic harmony, uniform sequences, and quantized voice leadings", *Music Theory Spectrum* **37**:1 (2015), 120–143.

berrycam@msu.edu                *Department of Mathematics, Michigan State University,*
                                *East Lansing, MI, United States*

tmfiore@umich.edu               *Department of Mathematics and Statistics,*
                                *University of Michigan-Dearborn, Dearborn, MI, United States*

                                *NWF I - Mathematik, Universität Regensburg, Regensburg,*
                                *Germany*

# On computable classes of equidistant sets: finite focal sets

Csaba Vincze, Adrienn Varga, Márk Oláh,
László Fórián and Sándor Lőrinc

(Communicated by Michael Dorff)

The equidistant set of two nonempty subsets $K$ and $L$ in the Euclidean plane is the set of all points that have the same distance from $K$ and $L$. Since the classical conics can be also given in this way, equidistant sets can be considered as one of their generalizations: $K$ and $L$ are called the focal sets. The points of an equidistant set are difficult to determine in general because there are no simple formulas to compute the distance between a point and a set. As a simplification of the general problem, we are going to investigate equidistant sets with finite focal sets. The main result is the characterization of the equidistant points in terms of computable constants and parametrization. The process is presented by a Maple algorithm. Its motivation is a kind of continuity property of equidistant sets. Therefore we can approximate the equidistant points of $K$ and $L$ with the equidistant points of finite subsets $K_n$ and $L_n$. Such an approximation can be applied to the computer simulation, as some examples show in the last section.

## 1. Introduction: notation and preliminaries

Let $K \subset \mathbb{R}^2$ be a subset in the Euclidean coordinate plane. The distance between a point $(x, y)$ and $K$ is measured by the usual infimum formula:

$$d((x, y), K) := \inf\{d((x, y), (a, b)) \mid (a, b) \in K\}.$$

Let us define the equidistant set of $K$ and $L \subset \mathbb{R}^2$ as the set of all points that have the same distance from $K$ and $L$:

$$\{K = L\} := \{(x, y) \in \mathbb{R}^2 \mid d((x, y), K) = d((x, y), L)\}.$$

The equidistant sets can be considered as a kind of generalization of conics [Ponce and Santibáñez 2014]: $K$ and $L$ are called the focal sets. Equidistant sets are often called midsets. Their investigations were started by Wilker [1975] and Loveland

[1976]. For another generalization of the classical conics and their applications, see, e.g., [Erdős and Vincze 1958; Melzak and Forsyth 1977] for polyellipses and their applications, and [Gross and Strempel 1998; Nagy and Vincze 2010; Vincze and Nagy 2011; 2012]. "We find equidistant sets as conventionally defined frontiers in territorial domain controversies: for instance, the United Nations Convention on the Law of the Sea (Article 15) establishes that, in absence of any previous agreement, the delimitation of the territorial sea between countries occurs exactly on the median line every point of which is equidistant of the nearest points to each country"; for the citation, see [Ponce and Santibáñez 2014].

Let $R > 0$ be a positive real number. The *parallel body* of a set $K \subset \mathbb{R}^2$ with radius $R$ is the union of the closed disks with radius $R$ centered at the points of $K$. The infimum of the positive numbers such that $L$ is a subset of the parallel body of $K$ with radius $R$ and vice versa is called the *Hausdorff distance* of $K$ and $L$. It is well known that the Hausdorff metric makes the family of nonempty closed and bounded (i.e., compact) subsets in the plane a complete metric space; for the details, see, e.g., [Lay 1982; Vincze 2013]. In what follows we are going to characterize the equidistant points of finite focal sets in terms of computable constants and parametrization. The process will be presented by a Maple algorithm. Its motivation is the continuity property of equidistant sets in the sense of the following theorem.

**Theorem 1** [Ponce and Santibáñez 2014, Theorem 11]. *If $K$ and $L$ are disjoint compact subsets in the plane, and $K_n \to K$ and $L_n \to L$ are convergent sequences of nonempty compact subsets with respect to the Hausdorff metric then for any $R > 0$ we have*

$$\{K_n = L_n\} \cap \bar{D}(R) \to \{K = L\} \cap \bar{D}(R),$$

*where $\bar{D}(R)$ denotes the closed disk with radius $R$ centered at the origin.*

Since any compact subset can be approximated by finite subsets with respect to the Hausdorff metric, we can approximate the equidistant points of $K$ and $L$ with the equidistant points of finite subsets $K_n$ and $L_n$. Such an approximation can be applied to the computer simulation as an alternative to the error estimation process for quasiequidistant points suggested by [Ponce and Santibáñez 2014, §4.2].

## 2. The main result

Let $K, L \subset \mathbb{R}^2$ be nonempty finite disjoint subsets in the Euclidean coordinate plane:

$$K := \{(a_i, b_i) \mid i = 1, \ldots, p\} \quad \text{and} \quad L := \{(c_k, d_k) \mid k = 1, \ldots, q\},$$

where $p$ and $q$ are positive integers. Since we have only finitely many lines determined by the points of $K \cup L$ we can use the following technical condition without loss of generality:

(H) Each line determined by the points of $K \cup L$ has a slope different from zero; i.e., there are no horizontal "focal lines".

Indeed, an infinitesimal rotation about the origin provides the configuration we need to satisfy condition (H). On the other hand, the inverse rotation takes the equidistant points of the rotated sets into the equidistant points of the original ones. Let

$$K_i := \left\{ (x, y) \in \mathbb{R}^2 \mid d((x, y), K) = d((x, y), (a_i, b_i)) \right\} \quad (i = 1, \dots, p).$$

It is clear that

$$K_i = \bigcap_{\substack{j=1,\dots,m \\ j \neq i}} F_{ij},$$

where the closed half-planes $F_{ij}$ ($i \neq j$) are determined by the perpendicular bisector

$$(a_i - a_j)x + (b_i - b_j)y = \frac{a_i^2 - a_j^2}{2} + \frac{b_i^2 - b_j^2}{2} \tag{1}$$

of the segment $(a_i, b_i)$ and $(a_j, b_j)$ such that $(a_i, b_i) \in F_{ij}$. For any index $i$, the set $K_i$ is closed and convex as the intersection of finitely many closed half-planes. It is nonempty because $(a_i, b_i) \in K_i$. Since $K_i \cap K_j$ ($i \neq j$) is a subset of the perpendicular bisector (1) of the corresponding focal points in $K$, we can conclude that $\operatorname{int} K_i \cap \operatorname{int} K_j = \varnothing$ for any $i \neq j$. Finally,

$$\mathbb{R}^2 = \bigcup_{i=1}^{p} K_i;$$

i.e., we have a partitioning of the plane into (nonempty, closed and convex) regions with pairwise disjoint interiors based on the distance to points in a specific subset. It is called the *Voronoi decomposition*.

**Exercise 1.** Prove that $K_i$ is bounded if and only if $(a_i, b_i)$ is in the interior of the convex hull of $K$.

The Voronoi decomposition of the plane with respect to the points of $K$ means that the plane is divided into (nonempty, closed and convex) regions with pairwise disjoint interiors such that the distance of $(x, y) \in K_i$ to the focal set $K$ can be measured as the distance of $(x, y) \in K_i$ to $(a_i, b_i) \in K$. In terms of inequalities,

$$K_i : (a_i - a_j)x + (b_i - b_j)y \geq \frac{a_i^2 - a_j^2}{2} + \frac{b_i^2 - b_j^2}{2}, \tag{2}$$

where $j$ runs from 1 to $p$ but $i \neq j$. Using condition (H) we can reformulate the system of inequalities as

$$\begin{aligned} K_i : y \geq \alpha_{ij}x + \beta_{ij} \quad & (b_i - b_j > 0), \\ y \leq \alpha_{ij}x + \beta_{ij} \quad & (b_i - b_j < 0), \end{aligned} \tag{3}$$

where

$$\alpha_{ij} = -\frac{a_i - a_j}{b_i - b_j}, \quad \beta_{ij} = \frac{1}{b_i - b_j}\left(\frac{a_i^2 - a_j^2}{2} + \frac{b_i^2 - b_j^2}{2}\right) \quad \text{and} \quad i \neq j.$$

In a similar way consider the Voronoi decomposition of the plane with respect to the points of $L$:

$$L_k := \left\{(x, y) \in \mathbb{R}^2 \mid d((x, y), L) = d((x, y), (c_k, d_k))\right\} \quad (k = 1, \ldots, q),$$

$$L_k = \bigcap_{\substack{l=1,\ldots,q \\ l \neq k}} F_{kl},$$

where the closed half-planes $F_{kl}$ $(k \neq l)$ are determined by the perpendicular bisector

$$(c_k - c_l)x + (d_k - d_l)y = \frac{c_k^2 - c_l^2}{2} + \frac{d_k^2 - d_l^2}{2} \tag{4}$$

of the segment $(c_k, d_k)$ and $(c_l, d_l)$ such that $(c_k, d_k) \in F_{kl}$,

$$\mathbb{R}^2 = \bigcup_{k=1}^{q} L_k.$$

In terms of inequalities,

$$L_k : (c_k - c_l)x + (d_k - d_l)y \geq \frac{c_k^2 - c_l^2}{2} + \frac{d_k^2 - d_l^2}{2}, \tag{5}$$

where $l$ runs from 1 to $q$ but $k \neq l$. Using condition (H) we can reformulate the system of inequalities as

$$\begin{aligned} L_k : y &\geq \gamma_{kl}x + \delta_{kl} \quad (d_k - d_l > 0), \\ y &\leq \gamma_{kl}x + \delta_{kl} \quad (d_k - d_l < 0), \end{aligned} \tag{6}$$

where

$$\gamma_{kl} = -\frac{c_k - c_l}{d_k - d_l}, \quad \delta_{kl} = \frac{1}{d_k - d_l}\left(\frac{c_k^2 - c_l^2}{2} + \frac{d_k^2 - d_l^2}{2}\right) \quad \text{and} \quad k \neq l.$$

**Lemma 1.** *The set of equidistant points is equal to the union of*

$$\bigcup_{i=1}^{p}\bigcup_{k=1}^{q}(K_i \cap L_k \cap l_{ik}),$$

*where*

$$l_{ik} : (a_i - c_k)x + (b_i - d_k)y = \frac{a_i^2 - c_k^2}{2} + \frac{b_i^2 - d_k^2}{2} \tag{7}$$

*is the perpendicular bisector of* $(a_i, b_i)$ *and* $(c_k, d_k)$.

In what follows we characterize the sets of the form $K_i \cap L_k \cap l_{ik}$ in terms of a system of linear inequalities. According to condition (H), equation (7) of the perpendicular bisector $l_{ik}$ can be written in the form

$$l_{ik} : y = \mu_{ik}x + \nu_{ik}, \tag{8}$$

where

$$\mu_{ik} := -\frac{a_i - c_k}{b_i - d_k} \quad \text{and} \quad \nu_{ik} = \frac{1}{b_i - d_k}\left(\frac{a_i^2 - c_k^2}{2} + \frac{b_i^2 - d_k^2}{2}\right). \tag{9}$$

This means by (3) and (6) that

$$
\begin{aligned}
K_i \cap L_k \cap l_{ik} : \mu_{ik}x + \nu_{ik} &\geq \alpha_{ij}x + \beta_{ij} \quad (b_i - b_j > 0), \\
\mu_{ik}x + \nu_{ik} &\leq \alpha_{ij}x + \beta_{ij} \quad (b_i - b_j < 0), \\
\mu_{ik}x + \nu_{ik} &\geq \gamma_{kl}x + \delta_{kl} \quad (d_k - d_l > 0), \\
\mu_{ik}x + \nu_{ik} &\leq \gamma_{kl}x + \delta_{kl} \quad (d_k - d_l < 0),
\end{aligned}
\tag{10}
$$

where $j$ runs from 1 to $p$ but $j \neq i$ and $l$ runs from 1 to $q$ but $l \neq k$. It can be easily seen that the number of inequalities is $p + q - 2$ for any fixed pair of indices $(i, k)$ and the equidistant set is the union of finitely many polygonal chains determined by inequalities of type (10). To reduce the number of possible cases, we formulate necessary and sufficient conditions for the solvability of system (10). Let us introduce the following set of indices:

$$
\begin{aligned}
P_{ik}^+ &:= \{j \mid (b_i - b_j)(\mu_{ik} - \alpha_{ij}) > 0\}, \\
P_{ik}^- &:= \{j \mid (b_i - b_j)(\mu_{ik} - \alpha_{ij}) < 0\}, \\
P_{ik}^{0+} &:= \{j \mid b_i - b_j > 0 \text{ and } \mu_{ik} - \alpha_{ij} = 0\}, \\
P_{ik}^{0-} &:= \{j \mid b_i - b_j < 0 \text{ and } \mu_{ik} - \alpha_{ij} = 0\}, \\
Q_{ik}^+ &:= \{l \mid (d_k - d_l)(\mu_{ik} - \gamma_{kl}) > 0\}, \\
Q_{ik}^- &:= \{l \mid (d_k - d_l)(\mu_{ik} - \gamma_{kl}) < 0\}, \\
Q_{ik}^{0+} &:= \{l \mid d_k - d_l > 0 \text{ and } \mu_{ik} - \gamma_{kl} = 0\}, \\
Q_{ik}^{0-} &:= \{l \mid d_k - d_l < 0 \text{ and } \mu_{ik} - \gamma_{kl} = 0\}.
\end{aligned}
\tag{11}
$$

Then we have that $(x, y) \in K_i \cap L_k \cap l_{ik}$ if and only if the following conditions are satisfied:

$$x \geq \frac{\beta_{ij} - \nu_{ik}}{\mu_{ik} - \alpha_{ij}} \quad (j \in P_{ik}^+) \qquad \text{and} \qquad x \geq \frac{\delta_{kl} - \nu_{ik}}{\mu_{ik} - \gamma_{kl}} \quad (l \in Q_{ik}^+), \tag{12}$$

$$x \leq \frac{\beta_{ij} - \nu_{ik}}{\mu_{ik} - \alpha_{ij}} \quad (j \in P_{ik}^-) \qquad \text{and} \qquad x \leq \frac{\delta_{kl} - \nu_{ik}}{\mu_{ik} - \gamma_{kl}} \quad (l \in Q_{ik}^-), \tag{13}$$

$$v_{ik} \geq \beta_{ij} \qquad (j \in P_{ik}^{0+}) \qquad \text{and} \qquad v_{ik} \geq \delta_{kl} \qquad (l \in Q_{ik}^{0+}), \qquad (14)$$

$$v_{ik} \leq \beta_{ij} \qquad (j \in P_{ik}^{0-}) \qquad \text{and} \qquad v_{ik} \leq \delta_{kl} \qquad (l \in Q_{ik}^{0-}). \qquad (15)$$

Therefore we can formulate the sufficient and necessary conditions in terms of the following constants:

$$m_{ik}^K := \begin{cases} -\infty & \text{if } P_{ik}^+ = \varnothing, \\ \sup_{j \in P_{ik}^+} (\beta_{ij} - v_{ik})/(\mu_{ik} - \alpha_{ij}) & \text{otherwise,} \end{cases} \qquad (16)$$

$$m_{ik}^L := \begin{cases} -\infty & \text{if } Q_{ik}^+ = \varnothing, \\ \sup_{l \in Q_{ik}^+} (\delta_{kl} - v_{ik})/(\mu_{ik} - \gamma_{kl}) & \text{otherwise,} \end{cases} \qquad (17)$$

$$M_{ik}^K := \begin{cases} \infty & \text{if } P_{ik}^- = \varnothing, \\ \inf_{j \in P_{ik}^-} (\beta_{ij} - v_{ik})/(\mu_{ik} - \alpha_{ij}) & \text{otherwise,} \end{cases} \qquad (18)$$

$$M_{ik}^L := \begin{cases} \infty & \text{if } Q_{ik}^- = \varnothing, \\ \inf_{l \in Q_{ik}^-} (\delta_{kl} - v_{ik})/(\mu_{ik} - \gamma_{kl}) & \text{otherwise,} \end{cases} \qquad (19)$$

$$r_{ik}^K := \begin{cases} -\infty & \text{if } P_{ik}^{0+} = \varnothing, \\ \sup_{j \in P_{ik}^{0+}} \beta_{ij} & \text{otherwise,} \end{cases} \qquad (20)$$

$$r_{ik}^L := \begin{cases} -\infty & \text{if } Q_{ik}^{0+} = \varnothing, \\ \sup_{l \in Q_{ik}^{0+}} \delta_{kl} & \text{otherwise,} \end{cases} \qquad (21)$$

$$R_{ik}^K := \begin{cases} \infty & \text{if } P_{ik}^{0-} = \varnothing, \\ \inf_{j \in P_{ik}^{0-}} \beta_{ij} & \text{otherwise,} \end{cases} \qquad (22)$$

$$R_{ik}^L := \begin{cases} \infty & \text{if } Q_{ik}^{0-} = \varnothing, \\ \inf_{l \in Q_{ik}^{0-}} \delta_{kl} & \text{otherwise,} \end{cases} \qquad (23)$$

$$m_{ik} := \sup\{m_{ik}^K, m_{ik}^L\}, \quad M_{ik} := \inf\{M_{ik}^K, M_{ik}^L\}, \qquad (24)$$

$$r_{ik} := \sup\{r_{ik}^K, r_{ik}^L\}, \quad R_{ik} := \inf\{R_{ik}^K, R_{ik}^L\}. \qquad (25)$$

**Theorem 2.** *If $K$ and $L$ are disjoint finite subsets satisfying condition* (H) *then for any pair $(i, k)$ of indices, $K_i \cap L_k$ contains equidistant points if and only if*

$$m_{ik} \leq M_{ik} \quad \text{and} \quad r_{ik} \leq v_{ik} \leq R_{ik}.$$

*The parametrization of the line segment of the equidistant points in $K_i \cap L_k$ is*

$$y = \mu_{ik}x + v_{ik} \quad (m_{ik} \leq x \leq M_{ik}).$$

*Proof.* It is clear that in the case where

$$m_{ik} \leq M_{ik} \quad \text{and} \quad r_{ik} \leq v_{ik} \leq R_{ik},$$

conditions (12)–(15) are satisfied for any $m_{ik} \leq x \leq M_{ik}$. $\qquad \square$

## 3. A Maple algorithm

Our algorithm, available in the online supplement, is implemented in Maple. The input data are the lists of $K$ and $L$ containing the points of the focal sets, respectively. $K[i][1]$ and $K[i][2]$ denote the coordinates of the $i$-th point in the focal set $K$ for each $i \in \{1, 2, \ldots, p\}$, and $L[k][1]$ and $L[k][2]$ denote the coordinates of the $k$-th point in the focal set $L$ for each $k \in \{1, 2, \ldots, q\}$. The main procedure `equidistant` creates a plot of the equidistant set with focal sets $K$ and $L$. The Maple command `LinearUnivariateSystem` produces the solution of system (10).

The procedure `equidistant_ini` takes the lists $K$ and $L$ as input and returns the following six objects as output:

- $a$ is a list containing the slopes $a[i][j] := \alpha_{ij}$ for each $i, j \in \{1, 2, \ldots, p\}$ and $i \neq j$, while $a[i][j] = 0$ when $i = j$.

- $g$ is a list containing the slopes $g[k][l] := \gamma_{kl}$ for each $k, l \in \{1, 2, \ldots, q\}$ and $k \neq l$, while $g[k][l] = 0$ when $k = l$.

- $b$ is a list containing the constants $b[i][j] := \beta_{ij}$ for each $i, j \in \{1, 2, \ldots, p\}$ and $i \neq j$, while $b[i][j] = 0$ when $i = j$.

- $d$ is a list containing the constants $d[k][l] := \delta_{kl}$ for each $k, l \in \{1, 2, \ldots, q\}$ and $k \neq l$, while $d[k][l] = 0$ when $k = l$.

- $m$ is a list containing the slopes $m[i][k] := \mu_{ik}$ for each $i \in \{1, 2, \ldots, p\}$ and $k \in \{1, 2, \ldots, q\}$.

- $n$ is a list containing the constants $n[i][k] := \nu_{ik}$ for each $i \in \{1, 2, \ldots, p\}$ and $k \in \{1, 2, \ldots, q\}$.

The procedure `xmaxmin` returns the maximal and the minimal values of the first coordinates of the points in the focal sets $K$ and $L$, respectively. They appear in the range option of the "plot" command.

The procedure `equidistant_system` creates the system of inequalities (10) for each $i \in \{1, 2, \ldots, p\}$ and $k \in \{1, 2, \ldots, q\}$. In the input data, $a, g, b, d, m, n$ are the objects created by `equidistant_ini`.

The procedure `inequality_range` defines the ranges for the next procedure, `equidistant_grafikon`. The equidistant set can be considered as the graph of a piecewise linear, continuous real function. The domain of such a function can be split into a finite number of disjoint intervals such that the function is linear over each interval. The procedure defines the endpoints of such intervals. The operands of the local variable $T$ containing the solution of a linear univariate system of inequalities are of the forms $c_1 < x$ and $x < c_2$. If $T$ has at least (and, consequently, exactly) two operands of the forms $c_1 < x$ and $x < c_2$, respectively, then we have both lower and upper bounds for the solution. Otherwise we have only a lower or
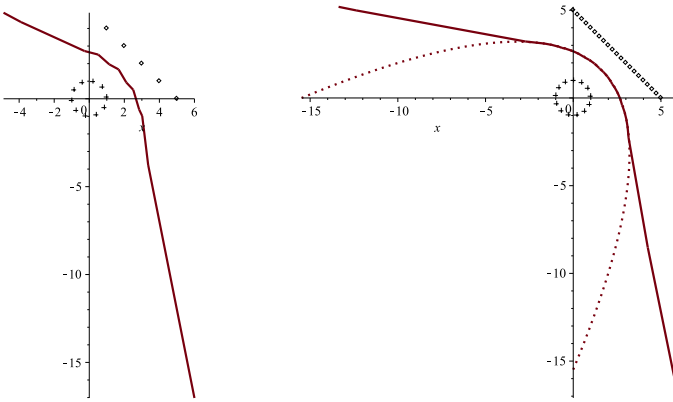
**Figure 1.** Examples 1 (left) and 2 (right).

an upper bound of the form $c_1 < x$ or $x < c_2$. In the case of $c_1 < x$, we choose the variable of numeric type as the lower bound for the range: $c_1$. The upper bound for the range is defined as the maximum of $c_1 + 1$ and xmax $+ 1$.

The procedure `equidistant_ grafikon` generates the list of plots of the graph of the linear functions, which represent the equidistant set with focal sets $K$ and $L$. In the input data, $m, n$ are created by `equidistant_ini` and $S$ is a list containing the list of ranges created by `inequality_range`.

**3.1. *Examples.*** We present some examples generated by the algorithm above. The code for Examples 1, 2, and 4 can be found in the online supplement.

**Example 1.** The focal set $K$ contains the points of a regular 10-gon inscribed in the unit circle; it is rotated by a small angle 0.1 to satisfy condition (H). The focal set $L$ contains the points $(1, 4)$, $(2, 3)$, $(3, 2)$, $(4, 1)$ and $(5, 0)$. They are lying on the same line segment $y = -x + 5$ $(0 \le x \le 5)$. See Figure 1.

**Example 2.** This case, shown in Figure 1, illustrates what happens when increasing the number of the focal points in Example 1.

The limit shape is a parabolic arc,

$$r(\varphi) = \frac{1 + 5/\sqrt{2}}{1 + \cos\left(\varphi - \frac{1}{4}\pi\right)}, \qquad (26)$$

provided that the polar angle belongs to the interval

$$-\arcsin \frac{12\sqrt{2}}{26 + 5\sqrt{2}} \le \varphi \le \frac{\pi}{2} + \arcsin \frac{12\sqrt{2}}{26 + 5\sqrt{2}}$$

because the line segment can be substituted by the entire line without changing the equidistancy in this region. Otherwise we have hyperbolic arcs because the distance to the line segment reduces to the distance from one of its endpoints.
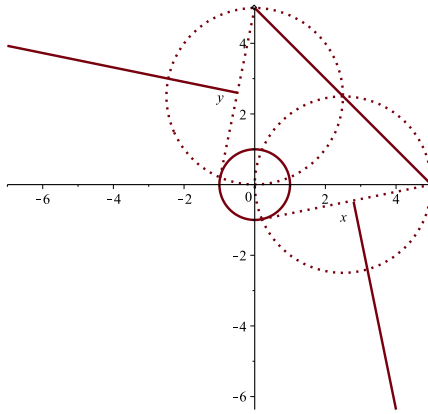
**Figure 2.** The asymptotic ends.

The asymptotic "ends", shown in Figure 2, are the bisectors of the points $P(0, 5)$, $T_2$ and $Q(5, 0)$, $T_4$, where $T_2$, $T_4$ denote the touching points of the tangent lines passing through $P$ and $Q$ in the second and the fourth quadrants, respectively. For the asymptotic behavior of the equidistant sets, see [Ponce and Santibáñez 2014, Theorem 12].

**Example 3.** The focal set $K$ contains the points $(-2, 1)$, $(-1, 1.3)$, $(0, 0)$, $(1, -2)$, $(2, -2.2)$, $(3, 1.5)$ and $(4, 3.5)$ and the focal set $L$ consists of $(1, 2)$, $(-1, 3)$, $(2, 4)$ and $(3, 5)$; see Figure 3.

**Example 4.** The focal set $K$ contains the points of a regular 7-gon inscribed in the circle of radius $\frac{1}{3}$ centered at the origin; it is also rotated by a small angle 0.1 to
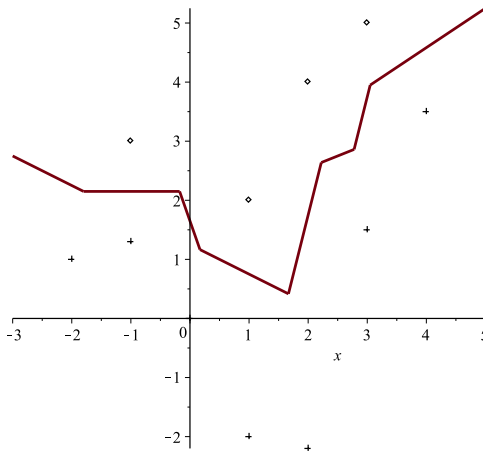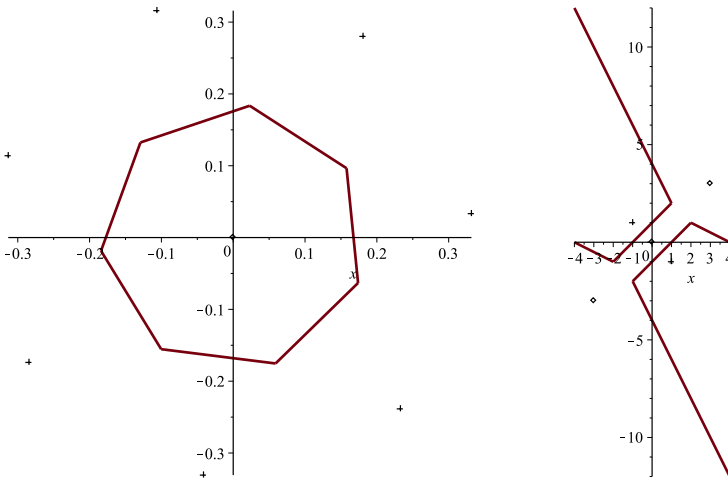


**Figure 3.** Example 3.

**Figure 4.** Examples 4 (left) and 5 (right).

satisfy condition (H). The focal set $L$ is a singleton containing the origin. In the same way any regular $n$-gon can be given as an equidistant set. See Figure 4.

**Example 5.** In this case, shown in Figure 4, we can see a disconnected case with focal sets $K$ containing the points $(-1, 1)$, $(1, -1)$ and $L$ containing the points $(-3, -3)$, $(0, 0)$, $(3, 3)$, respectively.

**3.2.** *Concluding remarks.* The application of the algorithm for more complicated focal sets is based on the continuity properties of the equidistant sets; see Section 1 and also [Ponce and Santibáñez 2014, Theorem 11].

Examples 1 and 2 represent the approximation of the equidistant set to a circle and a segment as focal sets. The Hausdorff distance can be estimated by comparing the polar distance of the equidistant points and the points of the limit shape as follows; see (26). First, get the polar coordinates of the vertex points of the approximating equidistant set. Since it is a polygonal chain, we have finitely many data depending on the number of the focal points: $(r_i, \varphi_i)$, where $i < \infty$. By (26) we can compute the exact polar distance $r(\varphi_i)$ belonging to the polar angle $\varphi_i$ on the limit parabola. Taking

$$D_1 := \max_i |r(\varphi_i) - r_i|,$$

we have an upper bound for the Hausdorff distance between the approximating equidistant set and the polygonal chain inscribed in the limit parabola with vertices of the polar angles $\varphi_i$. Indeed, if the polar body of a segment contains the endpoints of another one then it contains the entire line segment too. To estimate the Hausdorff distance of the inscribed polygonal chain and the parabolic arc, it is natural to consider the triangles $\triangle_i$ formed by adjacent vertices $V_i$ and $V_{i+1}$ of the polygonal

chain and the intersection of the tangent lines to the arc at $V_i$ and $V_{i+1}$. If $m_i$ denotes the height of the $\triangle_i$ belonging to the $i$-th side of the polygonal chain then its maximum $D_2$ gives an upper bound for the Hausdorff distance between the inscribed polygonal chain and the parabolic arc. Using the triangle inequality, the sum $D := D_1 + D_2$ is an upper bound for the Hausdorff distance between the estimating polygonal chain and the limit parabola.

In the same way, we can approximate the equidistant set to a pair of convex polytopes (the convex hulls of finite sets of points). Taking finitely many convex combinations of the vertices, one can produce finite focal sets to apply the algorithm. In case of general compact subsets we can use their intersections with a sequence of nested grids.

As the limit shape we have a circle by increasing the number of the vertices of the inscribed regular polygon (the focal set $K$) in Example 4. On the other hand, Example 4 shows a way of presenting regular polygons as equidistant sets. It has an important theoretical consequence in view of Weiszfeld's problem. E. Vázsonyi, also known as E. Weiszfeld, posed the problem of approximating convex plane curves with so-called polyellipses, all of whose points have the same sum of distances from finitely many focal points in the plane. It is the additive version of the approximation of plane curves by polynomial lemniscates, all of whose points have the same product of distances from finitely many focal points in the plane. P. Erdős and I. Vincze [1958] proved that the approximation of a regular triangle with polyellipses has an absolute error even if the number of focuses is increased to the infinity; see also [Varga and Vincze 2008]. This means that the idea of polyellipses gives an essentially different generalization of the classical conics.

## Acknowledgements

## References

[Erdős and Vincze 1958] P. Erdős and I. Vincze, "Über die Annäherung geschlossener, konvexer Kurven", *Mat. Lapok* **9** (1958), 19–36. MR Zbl

[Gross and Strempel 1998] C. Gross and T.-K. Strempel, "On generalizations of conics and on a generalization of the Fermat–Torricelli problem", *Amer. Math. Monthly* **105**:8 (1998), 732–743. MR Zbl

[Lay 1982] S. R. Lay, *Convex sets and their applications*, Wiley, New York, 1982. MR Zbl

[Loveland 1976] L. D. Loveland, "When midsets are manifolds", *Proc. Amer. Math. Soc.* **61**:2 (1976), 353–360. MR Zbl

[Melzak and Forsyth 1977] Z. A. Melzak and J. S. Forsyth, "Polyconics, I: Polyellipses and optimization", *Quart. Appl. Math.* **35**:2 (1977), 239–255. MR Zbl

[Nagy and Vincze 2010] Á. Nagy and Cs. Vincze, "Examples and notes on generalized conics and their applications", *Acta Math. Acad. Paedagog. Nyházi.* (*N.S.*) **26**:2 (2010), 359–375. MR Zbl

[Ponce and Santibáñez 2014] M. Ponce and P. Santibáñez, "On equidistant sets and generalized conics: the old and the new", *Amer. Math. Monthly* **121**:1 (2014), 18–32. MR Zbl

[Varga and Vincze 2008] A. Varga and Cs. Vincze, "On a lower and upper bound for the curvature of ellipses with more than two foci", *Expo. Math.* **26**:1 (2008), 55–77. MR Zbl

[Vincze 2013] Cs. Vincze, "Convex geometry", digital textbook, University of Debrecen, Hungary, 2013, available at http://www.tankonyvtar.hu/en/tartalom/tamop412A/2011_0025_mat_14/index.html.

[Vincze and Nagy 2011] Cs. Vincze and Á. Nagy, "An introduction to the theory of generalized conics and their applications", *J. Geom. Phys.* **61**:4 (2011), 815–828. MR Zbl

[Vincze and Nagy 2012] Cs. Vincze and Á. Nagy, "On the theory of generalized conics with applications in geometric tomography", *J. Approx. Theory* **164**:3 (2012), 371–390. MR Zbl

[Wilker 1975] J. B. Wilker, "Equidistant sets and their connectivity properties", *Proc. Amer. Math. Soc.* **47** (1975), 446–452. MR Zbl

csvincze@science.unideb.hu          *Institute of Mathematics, University of Debrecen, Hungary*

vargaa@eng.unideb.hu                *Faculty of Engineering, University of Debrecen, Hungary*

olma4000@gmail.com                  *BSc Mathematics, University of Debrecen, Hungary*

laci.forian@gmail.com               *BSc Mathematics, University of Debrecen, Hungary*

lorinc.sandor22@gmail.com           *BSc Electric Engineering, University of Debrecen, Hungary*

# Zero divisor graphs of commutative graded rings

Katherine Cooper and Brian Johnson

(Communicated by Scott T. Chapman)

We study a natural generalization of the zero divisor graph introduced by Anderson and Livingston to commutative rings graded by abelian groups, considering only homogeneous zero divisors. We develop a basic theory for graded zero divisor graphs and present many examples. Finally, we examine classes of graphs that are realizable as graded zero divisor graphs and close with some open questions.

## 1. Introduction

Zero divisor graphs of commutative rings have been well-studied since their introduction by Beck [1988], and there have also been many generalizations, from noncommutative rings to semigroups. Anderson and Livingston [1999] began studying the graph created from just the nonzero zero divisors. We focus on a generalization of their graph to graded rings. In this way we are able to realize significantly more graphs as graded zero divisor graphs. While the class of realizable graphs is expanded, some of the same restrictions still exist in the graded case. For other types of graphs associated to graded rings, see [Khosh-Ahang and Nazari-Moghadam 2016]. For examples of other graphs associated to commutative rings, see [Anderson and Badawi 2012; Ashrafi et al. 2010; Badawi 2014; 2015; Behboodi and Rakeei 2011]. For more examples in the commutative case and characterizations based on numbers of zero divisors, among other things, see [Anderson and Badawi 2008].

In Section 2 we summarize the basic notation, terminology, and necessary facts for graded rings. We also define the graded zero divisor graph and give some basic examples.

Section 3 contains the basic properties and theory of graded zero divisor graphs. As mentioned, many of the familiar properties from the nongraded case hold true in the graded case: the graded zero divisor graph is connected with diameter less than or equal to 3, the girth is less than or equal to 4 (when finite), and the graph is finite if and only if the ring is finite.

The final section is devoted to realizability of various graphs and classes of graphs. We show that all but one of the connected graphs on four vertices are realizable as graded zero divisor graphs, and we completely classify the connected graphs on five vertices. Further, we show that every star, complete, and complete bipartite graph is realizable, a marked difference from the nongraded case. We also include some interesting open questions.

Throughout the paper, all rings are assumed to be commutative with identity, and $G$ will always represent an abelian group.

## 2. Preliminaries

We now summarize some basic language and notation relating to rings graded by abelian groups as well as zero divisor graphs associated with such rings. For more details on graded commutative rings, the reader is referred to [Johnson 2012]. For a more general treatment, see [Năstăsescu and Van Oystaeyen 2004].

***Graded rings.*** Let $G$ be an abelian group. A *G-graded ring $R$* is a ring $R$ with a family of subgroups $\{R_g \mid g \in G\}$ of $R$ such that $R = \bigoplus_{g \in G} R_g$ (as abelian groups) and $R_g R_h \subseteq R_{g+h}$ for all $g, h \in G$. At times, we may refer simply to the "graded ring $R$" if $G$ is understood. If $r \in R$ then there exist unique elements $r_g \in R_g$ for each $g \in G$, all but finitely many of which are zero, such that $r = \sum_{g \in G} r_g$. If $r = r_g$ for some $g \in G$ then $r$ is called *G-homogeneous of degree g* (or simply "homogeneous"). An ideal $I \subseteq R$ is *G*-homogeneous (again, "homogeneous" when appropriate) provided $I = \bigoplus_{g \in G} I_g$ for some family of subgroups $\{I_g \mid g \in G\}$. Equivalently, we only need know that $I$ has a generating set consisting of homogeneous elements.

When defining some basic ring-theoretic properties in terms of only homogeneous elements, we incorporate the grading group to simplify language and avoid confusion. For example, a *G*-graded ring $R$ is called a *G-field* (respectively, *G-domain*) if every nonzero *G*-homogeneous element of $R$ is a unit (respectively, not a zero divisor). Note that when we refer to a property holding under the trivial grading, or 0-grading, we will not write "$R$ is a 0-field," but rather "$R$ is a field."

The following lemma is interesting on its own. It says that to decompose a graded ring as a (graded) direct product, it is enough to write the ring as a direct product of subrings. We use it later in our analysis of realizable graphs.

**Lemma 2.1.** *Suppose $R$ is a $G$-graded ring, and $R = S \times T$ for subrings $S$ and $T$ of $R$. Then $S$ and $T$ are $G$-graded subrings of $R$, and $R$ is the (graded) direct product of $S$ and $T$.*

*Proof.* As above, suppose $R = S \times T$. Define $S_g := \{s \in S \mid (s, 0) \in R_g\}$ and $T_g := \{t \in T \mid (0, t) \in R_g\}$. This defines a $G$-grading on $S$ and $T$, and so it only remains to be shown that $R$ is their graded direct product.

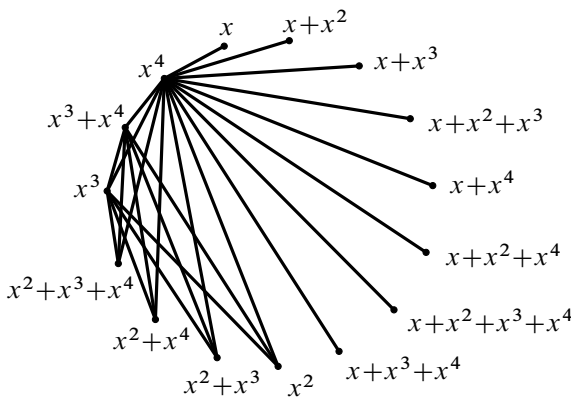By Remark 1.2.3 in [Năstăsescu and Van Oystaeyen 2004], we are done.    □

**Figure 1.** $\Gamma(R)$.

***Zero divisor graphs.*** Let $R$ be a $G$-graded ring, and let $Z_G^*(R)$ denote the collection of nonzero $G$-homogeneous zero divisors. Define the *$G$-graded zero divisor graph* (or just the "graded zero divisor graph" if $G$ is understood) $\Gamma_G(R)$ to be the graph whose vertices are the elements of $Z_G^*(R)$ and which has an edge between distinct elements $x, y \in Z_G^*(R)$ provided $xy = 0$. It is worth mentioning that one could eliminate the restriction that $x$ and $y$ be distinct; the only change is that the graphs now might have loops. However, the graph theory becomes significantly more complicated. See [Vietri 2015] for examples of classifications involving loops.

As in the case of 0-fields, for example, when we consider a trivial grading, we use $Z(R)$, $Z^*(R)$, and $\Gamma(R)$ rather than include the subscript 0.

One interesting result of studying a graded version of zero divisor graphs is that the same ring may have different gradings, leading to distinct graphs from the same underlying ring.

**Example 2.2.** Let $R = \mathbb{Z}_2[X]/(X^5)$ and use $x$ to denote the image of $X$ in the quotient.

(1) Consider $R$ under a trivial grading. That is, suppose the degree of every element is 0 (so $G$ could be any abelian group, in fact). Since all elements of $R$ are homogeneous, this is the same as the usual zero divisor graph $\Gamma(R)$, as shown in Figure 1.

(2) Now consider $R$ as a $\mathbb{Z}_2$-graded ring under the assignment induced by $\deg(x) = 1$, so the degree of $x^i$ is $i \pmod 2$. This restricts the number of homogeneous elements and homogeneous zero divisors, as shown in Figure 2. For example, $x^2 + x^4$ is homogeneous, but $x^2 + x^3$ is not.

(3) Finally, consider $R$ as a $\mathbb{Z}$-graded ring under the assignment induced by $\deg(x) = 1$, so the degree of $x^i$ is $i$. This further restricts the number of homogeneous zero divisors, as seen in Figure 3. In fact, the only homogeneous zero divisors are elements of the form $x^i$, for $i = 1, 2, 3, 4$.
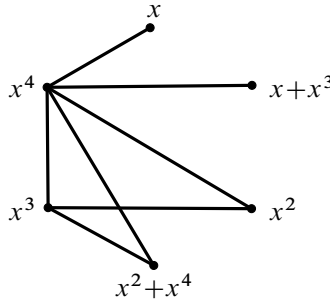
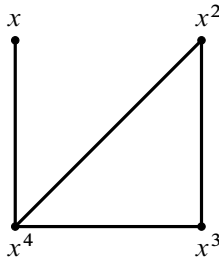**Figure 2.** $\Gamma_{\mathbb{Z}_2}(R)$.



**Figure 3.** $\Gamma_{\mathbb{Z}}(R)$.

It is worth mentioning that the gradings on the first two rings can be induced from the third ring. In general, given a $G$-graded ring $R$ and a subgroup $H$ of $G$, there is a natural grading of $R$ by the quotient $G/H$, obtained by setting $R_{g+H} = \bigoplus_{h \in H} R_{g+h}$. For instance, to obtain the $\mathbb{Z}_2$-grading of $R$ from the $\mathbb{Z}$-grading, we take $G = \mathbb{Z}$ and $H = 2\mathbb{Z}$, whereas to obtain the trivial grading, we take $G = H = \mathbb{Z}$.

## 3. Basic properties

Many of the basic properties of $\Gamma(R)$ described by Anderson and Livingston [1999] have analogues for $\Gamma_G(R)$. For example, they show that the zero divisor graph is finite if and only if $R$ is finite or a domain. With modifications we can use a similar proof, combined with the following lemma, to prove a corresponding result for graded rings.

**Lemma 3.1.** *If $Z_G(R)$ is finite, then for every $x \in Z_G^*(R)$, $\mathrm{ann}(x)$ is finite.*

*Proof.* Let $I = \mathrm{ann}(x)$. As $x$ is homogeneous, $I$ is homogeneous, and thus $I = \bigoplus_{g \in G} I_g$. Further, $I_g \subseteq Z_G(R)$ for every $g \in G$, so $I_g = 0$ for all but finitely many $g \in G$ and each nonzero $I_g$ is finite. Since there are finitely many nonzero $I_g$, say $I_{g_1}, \ldots, I_{g_k}$, we have $|I| = \left| \bigoplus_{i=1}^{k} I_{g_i} \right| = \prod_{i=1}^{k} |I_{g_i}|$. Therefore $|I| < \infty$. $\square$

**Theorem 3.2.** *Let $R$ be a commutative ring. Then $|\Gamma_G(R)|$ is finite if and only if $R$ is a $G$-domain or $R$ is finite.*

*Proof.* Suppose $R$ is not a $G$-domain and $|Z_G^*(R)|$ is finite. Then there exist nonzero homogeneous $x, y \in R$ with $xy = 0$. Let $I = \text{ann}(x)$. By Lemma 3.1, $I$ is finite. Also, $ry \in I$ for all $r \in R$. If $R$ is infinite, then there exists $i \in I$ with $B = \{r \in R \mid ry = i\}$ infinite. For any $r, s \in B$, we have $(r - s)y = 0$, so $\text{ann}(y)$ is infinite, contradicting Lemma 3.1. Thus $R$ must be finite. $\square$

Because there is no "graded" version of the ring being finite, we get an interesting corollary.

**Corollary 3.3.** *If $1 \leq |Z_G^*(R)| < \infty$, then $1 \leq |Z^*(R)| < \infty$.*

*Proof.* Suppose $1 \leq |Z_G^*(R)| < \infty$. If $|Z^*(R)| = \infty$, then $R$ is not finite. Therefore, $R$ must be a $G$-domain, so $|Z_G^*(R)| = 0$, a contradiction. If $|Z_G^*(R)| \geq 1$, clearly $|Z^*(R)| \geq 1$. $\square$

**Note.** The converse of Corollary 3.3 is also true for the upper bounds, but fails when the lower bound 1 is added, as the following example shows.

**Example 3.4.** Consider

$$R := \frac{\mathbb{Z}_3[X]}{(X^2 - 1)} = \mathbb{Z}_3 \oplus \mathbb{Z}_3 x,$$

where $x$ is the image of $X$ in the quotient ring. This has a natural grading by $\mathbb{Z}_2$, where $\deg(x^i) = i \pmod 2$. One easily verifies that this grading makes $R$ a $\mathbb{Z}_2$-field. However, $(x + 1)(x - 1) = x^2 - 1 = 0$, so $|Z_G^*(R)| = 0$, yet $|Z^*(R)| \geq 1$.

Another obvious consequence of the finiteness result above is that we can assume a ring with a finite graded zero divisor graph is graded by a finitely generated group. Moreover, it can be shown that the grading group can be chosen to be finite. For example, if such a ring $R$ is graded by $\mathbb{Z}$, say $R = \bigoplus_{n \in \mathbb{Z}} R_n$, we can form the quotient group $G = \mathbb{Z}/k\mathbb{Z}$, where $k = \max\{m - \ell \mid R_m \neq 0 \text{ and } R_\ell \neq 0\}$. This argument can be extended to any finitely generated group by applying it in each component of the free part of the grading group as necessary.

Other well known facts about zero divisor graphs concern connectedness, diameter, and girth. None of these theorems change in the graded setting.

**Theorem 3.5.** *Let $G$ be an abelian group and $R$ a $G$-graded ring. Then $\Gamma_G(R)$ is connected and $\text{diam}(\Gamma_G(R)) \leq 3$.*

*Proof.* The proof given in [Anderson and Livingston 1999, Theorem 2.3] can be used if one simply adds that each zero divisor chosen is homogeneous. $\square$

Similarly, the following well-known result can be obtained by modifying the proof given by Axtell, Coykendall, and Stickles [Axtell et al. 2005], insisting that each choice of a zero divisor is homogeneous.

**Theorem 3.6.** *Suppose G is an abelian group and R is a G-graded ring. If $\Gamma_G(R)$ contains a cycle, then the girth of $\Gamma_G(R)$ is less than or equal to 4.*

Some of the previous facts can also be obtained by results on zero divisor graphs of semigroups found in [DeMeyer et al. 2002]. Indeed, the homogeneous elements (together with 0) in a ring are closed under the ring multiplication.

## 4. Realizability of Graphs

There has been ample study on which graphs are realizable as zero divisor graphs of commutative rings; for example, see [Axtell et al. 2009; LaGrange 2008; Redmond 2007]. Certainly, any graph realizable as $\Gamma(R)$ for a ring $R$ is realizable as $\Gamma_G(R)$ for the same ring under a trivial grading (by any group $G$). It turns out that there are significantly more graphs realizable as graded zero divisor graphs. We begin with graphs on four vertices, but every connected graph on one, two, or three vertices is realizable as the (nongraded) zero divisor graph of a commutative ring. Therefore, there is nothing to show in the graded case for these.

***Connected graphs on four vertices.*** Anderson and Livingston [1999] indicate that of the six connected graphs on four vertices, only those shown in Figure 4 may be realized as $\Gamma(R)$. Their proofs that the other three graphs seen in Figure 5 are not realizable all have a similar flavor. One uses the fact that certain sums or products must be annihilated by another element in the graph, and therefore must also be vertices in the zero divisor graph. This breaks down (often) in the graded case. Even though all of the vertices represent homogeneous elements and the sum of elements may still be annihilated, unless we know that both (homogeneous) elements are of *the same degree*, this sum no longer needs to be another vertex in the graded zero divisor graph.

For zero divisor graphs of graded rings, the three graphs in Figure 4 are still realizable, but we can also produce two more.

The graph on the left in Figure 5 is realized using the ring $\mathbb{Z}_2[X, Y]/(XY, X^2, Y^4)$ under the $\mathbb{Z}_2 \oplus \mathbb{Z}_4$-grading defined by $\deg(x) = (1 \ (\mathrm{mod}\ 2), 0 \ (\mathrm{mod}\ 4))$ and $\deg(y) = (0 \ (\mathrm{mod}\ 2), 1 \ (\mathrm{mod}\ 4))$, where $x$ and $y$ represent the images of $X$ and $Y$ in the quotient.
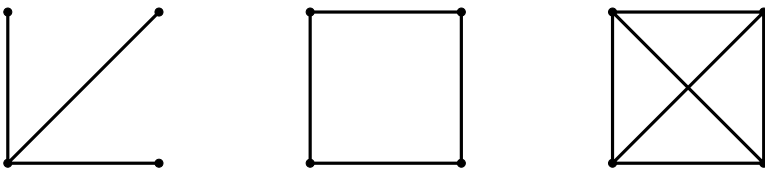


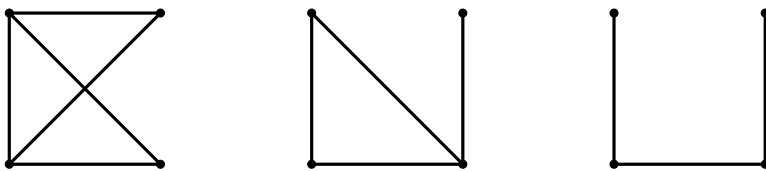**Figure 4.** The three connected graphs on four vertices realizable as $\Gamma(R)$.

**Figure 5.** Two additional graphs realizable as $\Gamma_G(R)$ (left, middle) and an unrealizable (right) connected graph on four vertices.

The graph in the middle is realized with the ring $\mathbb{Z}_2[X]/(X^5)$ under the $\mathbb{Z}$-grading defined by $\deg(x) = 1$, where $x$ is the image of $X$ in the quotient. We could also obtain the same graph using a $\mathbb{Z}_5$-grading and setting $\deg(x^i) = i \pmod 5$.

The final graph on the right in Figure 5 remains unrealizable as $\Gamma_G(R)$ for any group $G$. It can be proven that each of the four zero divisors must be (homogeneous) of the same degree, and thus the proof provided by Anderson and Livingston can be used.

***Connected graphs on five vertices.*** An interesting fact is that while there are 21 connected graphs on five vertices, there are still only three of these graphs realizable as $\Gamma(R)$. This can be proved using a mix of results from [Anderson and Livingston 1999] and direct analysis of adding and/or multiplying certain zero divisors together to reach a contradiction; alternatively, this is shown in [Redmond 2003]. These three graphs and the rings used to construct them are shown in Figure 6. Here, $\mathbb{F}_4$ represents a finite field with four elements.

As before, we are able to construct more of these graphs in the graded setting (in addition to those in Figure 6). Figure 7 summarizes the additional graphs we are able to realize, while Table 1 summarizes the grading used on each ring. In the table we use $x$ and $y$ to denote the images of $X$ and $Y$ in factor rings, while $e_i$ denotes the $i$-th basis vector, which has a 1 $\pmod n$ (for the appropriate $n$) in the $i$-th position and 0s elsewhere.

Not every connected graph on five vertices is realizable as a graded zero divisor graph. Figure 8 contains the graphs unrealizable as graded zero divisor graphs.
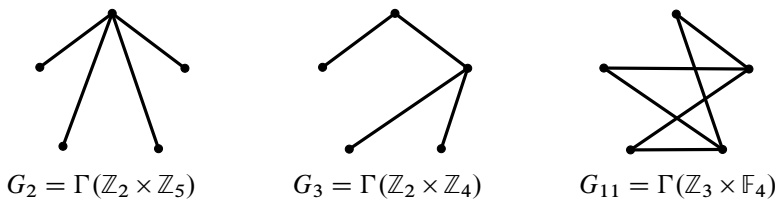


$$G_2 = \Gamma(\mathbb{Z}_2 \times \mathbb{Z}_5) \qquad G_3 = \Gamma(\mathbb{Z}_2 \times \mathbb{Z}_4) \qquad G_{11} = \Gamma(\mathbb{Z}_3 \times \mathbb{F}_4)$$

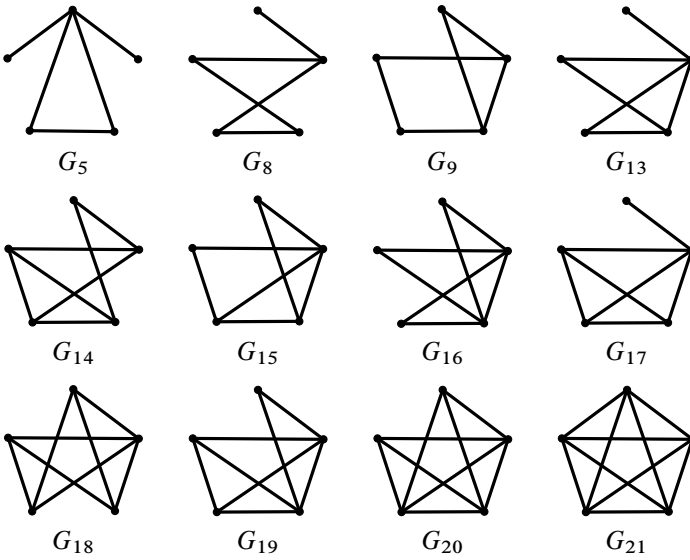**Figure 6.** Connected graphs on five vertices realizable as (non-graded) zero divisor graphs.

**Figure 7.** Additional connected graphs on five vertices realizable as graded zero divisor graphs.

| graph | ring | group | grading |
|-------|------|-------|---------|
| $G_5$ | $\frac{\mathbb{Z}_3[X]}{(X^2)} \times \mathbb{Z}_2$ | $\mathbb{Z}_2$ | $\deg((x,0)) = 1 \pmod 2$ |
| $G_8$ | $\frac{\mathbb{Z}_2[X]}{((X+1)^2 X^2)}$ | $\mathbb{Z}_2$ | $\deg(x^i) = i \pmod 2$ |
| $G_9$ | $\frac{\mathbb{Z}_2[X]}{(X^2)} \times \frac{\mathbb{Z}_2[Y]}{(Y^2)}$ | $\mathbb{Z}_2$ | $\deg((x,0)) = \deg((0,y)) = 1 \pmod 2$ |
| $G_{13}$ | $\frac{\mathbb{Z}_2[X]}{(X^6)}$ | $\mathbb{Z}_6$ | $\deg(x) = 1 \pmod 6$ |
| $G_{14}$ | $\frac{\mathbb{Z}_2[X]}{(X^3)} \times \mathbb{Z}_3$ | $\mathbb{Z}_3$ | $\deg((x,0)) = 1 \pmod 3$ |
| $G_{15}$ | $\frac{\mathbb{Z}_2[X,Y]}{(X^3,Y^2)}$ | $\mathbb{Z}_3 \oplus \mathbb{Z}_2$ | $\deg(x) = e_1,\ \deg(y) = e_2$ |
| $G_{16}$ | $\frac{\mathbb{Z}_2[X,Y]}{(XY,X^2,Y^4)}$ | $\mathbb{Z}_4$ | $\deg(x) = \deg(y) = 1 \pmod 4$ |
| $G_{17}$ | $\frac{\mathbb{Z}_2[X,Y]}{(X,Y)^2} \times \mathbb{Z}_2$ | $\mathbb{Z}_2$ | $\deg((x,0)) = \deg((y,0)) = 1 \pmod 2$ |
| $G_{18}$ | $\frac{\mathbb{Z}_2[X,Y]}{(XY,X^3-Y^3)}$ | $\mathbb{Z}_3 \oplus \mathbb{Z}_3$ | $\deg(x) = e_1,\ \deg(y) = e_2$ |
| $G_{19}$ | $\frac{\mathbb{Z}_2[X,Y]}{(XY^2,X^2,Y^4)}$ | $\mathbb{Z}_2 \oplus \mathbb{Z}_4$ | $\deg(x) = e_1,\ \deg(y) = e_2$ |
| $G_{20}$ | $\frac{\mathbb{Z}_2[X,Y]}{(XY,X^3,Y^3)}$ | $\mathbb{Z}_3$ | $\deg(x) = 1 \pmod 3,\ \deg(y) = 0$ |
| $G_{21}$ | $\frac{\mathbb{Z}_2[X_1,X_2,...,X_5]}{(X_i X_j \mid i,j \in \{1,2,...,5\})}$ | $(\mathbb{Z}_2)^5$ | $\deg(x_i) = e_i$ |

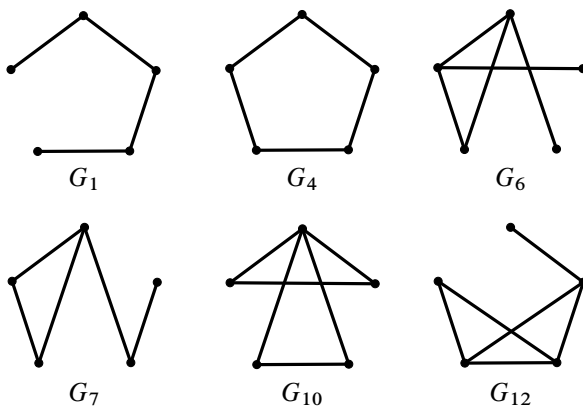**Table 1.** Rings and their gradings used to construct the graphs in Figure 7.

**Figure 8.** Connected graphs on five vertices unrealizable as a graded zero divisor graph.

Some can be eliminated easily, based on girth or diameter considerations, such as $G_1$ and $G_4$. To eliminate others, we used techniques similar to the nongraded case, with some modifications. To indicate the complications that arise, we provide an example.

**Example 4.1.** To show the graph $G_{10}$ is unrealizable, label the vertices $a$, $b$, $c$, $d$, and $e$ so that $a$ is the vertex at the top, continuing in alphabetical order clockwise.

From relations in the graph, we get that $bc$, $bd$, $ce$, and $de$ must be (nonzero) zero divisors. It is easily shown that each of these products must be equal to $a$. This implies $b, e \in R_g$ and $c, d \in R_h$ for some $g, h \in G$; that is, these elements are homogeneous of the same degree. Clearly, $b - e \in R_g$ and $b - e \neq 0$. Similarly, $c - d \in R_h$ and $c - d \neq 0$. As each of these differences is annihilated by $a$, we have $b - e, c - d \in Z_G^*(R)$.

We now simply exhaust all possibilities for $b - e$ and $c - d$. If $b - e = b$, then $e = 0$, a contradiction. If $b - e = e$, then $cb - ce = ce$, so that $a - a = a$, a contradiction. Similarly, we reach contradictions if $c - d \in \{c, d\}$. This gives $b - e \in \{a, c, d\}$ and $c - d \in \{a, b, e\}$.

Suppose $b - e = a$. Then $b, e, a \in R_g$. Thus, if $c - d \in \{a, b, e\}$, then $c, d \in R_g$, and the following statement holds:

(†) All five vertices are of the same degree, and $de = a$ (for example) implies this is degree 0. This implies $\Gamma_G(R) = \Gamma(R_0)$, but we know this graph cannot be realized as the usual zero divisor graph of any ring.

Now suppose $b - e = c$. Then $b, e, c, d \in R_g$. If $c - d = a$, then (†) applies again. If $c - d = b$, then $c - d - e = c$, so $d = -e$. This contradicts (for example) the fact that $ce \neq 0$. We obtain a similar contradiction if $c - d = e$.

Finally, suppose $b - e = d$. Then $b, e, c, d \in R_g$. Again, if $c - d = a$, (†) applies. If $c - d = b$, then $c - d - e = d$, so $bc - bd - be = bd$ gives us $a = 0$,

a contradiction. If $c - d = e$, then $b - c + d = d$, so $b = c$, a contradiction. It follows that $G_{10}$ cannot be realized as $\Gamma_G(R)$.

***Complete graphs.*** A central result of Anderson and Livingston [1999, Theorem 2.5] in their classification of realizable complete graphs (and in their classification of realizable star graphs, in fact) states that $\Gamma(R)$ has a vertex adjacent to every other vertex if and only if $R \cong \mathbb{Z}_2 \times A$, where $A$ is an integral domain, or $Z(R)$ is an annihilator ideal (and hence is prime). We prove a similar result in Theorem 4.3, using the following lemma.

**Lemma 4.2.** *Suppose $R$ is a $G$-graded ring and $a \in R$ is homogeneous. If $\mathrm{ann}(a)$ is maximal among annihilators of homogeneous elements, then $\mathrm{ann}(a)$ is $G$-prime.*

*Proof.* Suppose $x$ and $y$ are homogeneous and $xy \in \mathrm{ann}(a)$, but $x \notin \mathrm{ann}(a)$. We have $xa \neq 0$, but $xya = 0$. Thus $y \in \mathrm{ann}(xa)$. However, $\mathrm{ann}(xa) \subseteq \mathrm{ann}(a)$ implies $\mathrm{ann}(xa) = \mathrm{ann}(a)$. This implies $y \in \mathrm{ann}(a)$, and thus $\mathrm{ann}(a)$ is a $G$-prime ideal. $\square$

Because $Z_G(R)$ is very often not an ideal in the graded setting, we will end up considering $(Z_G(R))$, the ideal generated by the homogeneous zero divisors, in the theorem below.

**Theorem 4.3.** *Suppose $R$ is a $G$-graded ring. Then there is a vertex of $\Gamma_G(R)$ adjacent to every other vertex if and only if $R \cong \mathbb{Z}_2 \times A$, where $\mathbb{Z}_2$ and $A$ are $G$-graded and $A$ is a $G$-domain, or $(Z_G(R)) = \mathrm{ann}(x)$ for some nonzero homogeneous $x \in R$.*

*Proof.* ($\Leftarrow$) If $(Z_G(R)) = \mathrm{ann}(x)$, then $x$ is adjacent to every other vertex. If $R \cong \mathbb{Z}_2 \times A$, where $A$ is a $G$-domain, then $(1, 0)$ is adjacent to everything in $Z_G^*(R)$, except $(1, 0)$.

($\Rightarrow$) Suppose $(Z_G(R)) \neq \mathrm{ann}(x)$ for all nonzero homogeneous $x \in R$. Also, suppose there exists $a$ such that $0 \neq a \in Z_G(R)$ with $a$ adjacent to every other vertex.

If $a \in \mathrm{ann}(a)$, then $ax = 0$ for all $x \in Z_G(R)$. This implies $(Z_G(R)) \subseteq \mathrm{ann}(a)$. Also, $\mathrm{ann}(a)$ is homogeneous, so every homogeneous generator of $\mathrm{ann}(a)$ is in $Z_G(R)$. Thus $\mathrm{ann}(a) \subseteq (Z_G(R))$. So $\mathrm{ann}(a) = (Z_G(R))$, a contradiction. Therefore $a \notin \mathrm{ann}(a)$.

We claim $\mathrm{ann}(a)$ is maximal among those $\mathrm{ann}(x)$ such that $x$ is homogeneous. To see this, note that $a$ is adjacent to every other homogeneous zero divisor, yet $a \notin \mathrm{ann}(a)$.

By Lemma 4.2, $\mathrm{ann}(a)$ is $G$-prime. Since $a$ is a zero divisor, $a^2$ is also a homogeneous zero divisor. But $a \notin \mathrm{ann}(a)$, so $a^2 \neq 0$. If $a^2 \neq a$, then $a^2 \in \mathrm{ann}(a)$, but $\mathrm{ann}(a)$ is $G$-prime, so $a \in \mathrm{ann}(a)$, a contradiction. Therefore $a^2 = a$; that is, $a$ is a nontrivial (homogeneous) idempotent of degree 0.

By Lemma 2.1, $R = S \times T$ (as graded rings). Without loss of generality, let $a = (1, 0)$. Then $R = \mathbb{Z}_2 \times A$, where $A$ is a $G$-domain. $\square$

As we have seen in the examples above, we can construct graded zero divisor graphs that are complete for both four and five vertices. This already contrasts with the nongraded case, as Anderson and Livingston [1999, Theorem 2.10] show that only complete graphs on $p^n - 1$ vertices, where $p$ is prime and $n \geq 1$, are realizable as the zero divisor graph of a ring. In fact, in the graded case, we can realize every complete graph as a graded zero divisor graph. While we assume the graph is finite, the proof can easily be extended to infinite complete graphs.

**Theorem 4.4.** *A complete graph of any size is realizable as $\Gamma_G(R)$ for some abelian group $G$ and $G$-graded ring $R$.*

*Proof.* Consider $K_n$, the complete graph on $n$ vertices, where $n \geq 1$. Define the ring $S$ to be $\mathbb{Z}_2[X_1, \ldots, X_n]$, where the $X_i$ are indeterminates. This has an obvious grading by the group $G := \mathbb{Z}^n$, where we define the degree of $X_i$ to be $e_i$, the $i$-th basis vector in $G$ (which has a 1 in the $i$-th position and 0s elsewhere).

Let $I = (X_i X_j \mid i, j \in \{1, \ldots, n\})$ be the ideal generated by all products of two (not necessarily distinct) variables. As each generator is homogeneous, $I$ is a homogeneous ideal, and $R := S/I$ is also a $G$-graded ring.

One can now verify that $\Gamma_G(R) = K_n$ by noting that the only homogeneous elements in $R$ are the images of the $X_i$, all of which annihilate each other. $\square$

***Star graphs and complete bipartite graphs.*** Another well-studied class of graphs is the class of star graphs. A *star graph* is the complete bipartite graph $K_{1,k}$ for some $k \geq 0$. Except for the case $k = 0$, it can be thought of as having one vertex adjacent to all other vertices with no additional edges. Anderson and Livingston [1999, Theorem 2.13] completely characterized which star graphs are realizable for finite commutative rings. Star graphs were also studied by Coykendall, Sather-Wagstaff, Sheppardson, and Spiroff [Coykendall et al. 2012], but they focused on a different construction introduced by Mulay [2002], based on equivalence classes of zero divisors, denoted by $\Gamma_E(R)$.

For nongraded rings, it is only possible to realize the star graphs with $p^n$ vertices, where $p$ is a prime and $n \geq 0$. As with complete graphs, we can construct all (finite) star graphs in the graded setting. The following theorem is an obvious corollary of Theorem 4.6, and we omit the proof.

**Theorem 4.5.** *A star graph of any (finite) size is realizable as $\Gamma_G(R)$ for some abelian group $G$ and $G$-graded ring $R$.*

Not only can we realize all star graphs as graded zero divisor graphs, we can also realize *every* complete bipartite graph.

**Theorem 4.6.** *A complete bipartite graph of any (finite) size is realizable as $\Gamma_G(R)$ for some abelian group $G$ and $G$-graded ring $R$.*

*Proof.* Consider the graph $K_{m,n}$ and the rings defined by $S = \mathbb{Z}_2[X]/(X^m - 1)$ and $T = \mathbb{Z}_2[Y]/(Y^n - 1)$. Use $x$ and $y$, respectively, to denote the images of $X$ and $Y$ in $S$ and $T$. Define $L = \text{lcm}(m, n)$. Set $G = \mathbb{Z}_L$ and define $\mathbb{Z}_L$-gradings on $S$ and $T$, respectively, by setting $\deg(x) = \frac{L}{m}$ and $\deg(y) = \frac{L}{n}$. It is a straightforward exercise to show that each of these rings is now a $\mathbb{Z}_L$-field under its respective grading.

Form the graded direct product $R := S \times T$ (where $R_i = S_i \times T_i$). Notice that every nonzero element of $R$ of the form $(s, 0)$ or $(0, t)$, where $s \in S$ and $t \in T$ are homogeneous, is a vertex in $\Gamma_G(R)$. Also, each such element $(s, 0)$ is adjacent to each element $(0, t)$. Further, we claim these are the only vertices and edges in $\Gamma_G(R)$. To see this, suppose $(s_1, t_1)$ and $(s_2, t_2)$ are two elements of $Z_G^*(R)$. Because $S$ and $T$ are $\mathbb{Z}_L$-fields, and the $s_i$ and $t_i$ must be homogeneous, we can only have

$$(s_1, t_1)(s_2, t_2) = (0, 0)$$

when the elements on the left are of the form $(s_1, 0)$ and $(0, t_2)$ or $(0, t_1)$ and $(s_2, 0)$. □

## *Open questions.*

**Question 4.7.** Notice that for the constructions above, each ring is graded by a different abelian group. Another interesting question to consider is whether this is necessary. For example, for a fixed group $G$, can we still realize all complete graphs? If not, which graphs can we realize for a specific group?

**Question 4.8.** Theorem 4.3 is a step toward characterizing the graded rings that give rise to graded zero divisor graphs that are stars or complete graphs. A further avenue of study would be to determine if one can classify, completely or in part, the (graded) rings that give rise to star and/or complete graphs.

**Question 4.9.** Is there a generalization, in part or whole, of Theorem 4.6 to $n$-partite graphs? For example, Akbari, Maimani, and Yassemi [Akbari et al. 2003, Theorem 3.1] determine the rings whose zero divisor graphs are $n$-partite. They show, in particular, that if $n \geq 3$, at most one partitioning subset of $\Gamma(R)$ can have more than one vertex. As a contrast, graph $G_{18}$ in Figure 7 shows that in the graded case we can construct a complete 3-partite graph with more than one partitioning subset having size greater than 1.

## References

[Akbari et al. 2003] S. Akbari, H. R. Maimani, and S. Yassemi, "When a zero-divisor graph is planar or a complete $r$-partite graph", *J. Algebra* **270**:1 (2003), 169–180. MR Zbl

[Anderson and Badawi 2008] D. F. Anderson and A. Badawi, "On the zero-divisor graph of a ring", *Comm. Algebra* **36**:8 (2008), 3073–3092. MR Zbl

[Anderson and Badawi 2012]  D. F. Anderson and A. Badawi, "On the total graph of a commutative ring without the zero element", *J. Algebra Appl.* **11**:4 (2012), art. id. 1250074.  MR  Zbl

[Anderson and Livingston 1999]  D. F. Anderson and P. S. Livingston, "The zero-divisor graph of a commutative ring", *J. Algebra* **217**:2 (1999), 434–447.  MR  Zbl

[Ashrafi et al. 2010]  N. Ashrafi, H. R. Maimani, M. R. Pournaki, and S. Yassemi, "Unit graphs associated with rings", *Comm. Algebra* **38**:8 (2010), 2851–2871.  MR  Zbl

[Axtell et al. 2005]  M. Axtell, J. Coykendall, and J. Stickles, "Zero-divisor graphs of polynomials and power series over commutative rings", *Comm. Algebra* **33**:6 (2005), 2043–2050.  MR  Zbl

[Axtell et al. 2009]  M. Axtell, J. Stickles, and W. Trampbachls, "Zero-divisor ideals and realizable zero-divisor graphs", *Involve* **2**:1 (2009), 17–27.  MR  Zbl

[Badawi 2014]  A. Badawi, "On the annihilator graph of a commutative ring", *Comm. Algebra* **42**:1 (2014), 108–121.  MR  Zbl

[Badawi 2015]  A. Badawi, "On the dot product graph of a commutative ring", *Comm. Algebra* **43**:1 (2015), 43–50.  MR  Zbl

[Beck 1988]  I. Beck, "Coloring of commutative rings", *J. Algebra* **116**:1 (1988), 208–226.  MR  Zbl

[Behboodi and Rakeei 2011]  M. Behboodi and Z. Rakeei, "The annihilating-ideal graph of commutative rings, I", *J. Algebra Appl.* **10**:4 (2011), 727–739.  MR  Zbl

[Coykendall et al. 2012]  J. Coykendall, S. Sather-Wagstaff, L. Sheppardson, and S. Spiroff, "On zero divisor graphs", pp. 241–299 in *Progress in commutative algebra, II*, edited by C. Francisco et al., de Gruyter, Berlin, 2012.  MR  Zbl

[DeMeyer et al. 2002]  F. R. DeMeyer, T. McKenzie, and K. Schneider, "The zero-divisor graph of a commutative semigroup", *Semigroup Forum* **65**:2 (2002), 206–214.  MR  Zbl

[Johnson 2012]  B. P. Johnson, *Commutative rings graded by abelian groups*, Ph.D. thesis, University of Nebraska–Lincoln, 2012, available at http://search.proquest.com/docview/1038955241.

[Khosh-Ahang and Nazari-Moghadam 2016]  F. Khosh-Ahang and S. Nazari-Moghadam, "An associated graph to a graded ring", *Publ. Math. Debrecen* **88**:3-4 (2016), 401–416.  MR  Zbl

[LaGrange 2008]  J. D. LaGrange, "On realizing zero-divisor graphs", *Comm. Algebra* **36**:12 (2008), 4509–4520.  MR  Zbl

[Mulay 2002]  S. B. Mulay, "Cycles and symmetries of zero-divisors", *Comm. Algebra* **30**:7 (2002), 3533–3558.  MR  Zbl

[Năstăsescu and Van Oystaeyen 2004]  C. Năstăsescu and F. Van Oystaeyen, *Methods of graded rings*, Lecture Notes in Mathematics **1836**, Springer, 2004.  MR  Zbl

[Redmond 2003]  S. P. Redmond, "An ideal-based zero-divisor graph of a commutative ring", *Comm. Algebra* **31**:9 (2003), 4425–4443.  MR  Zbl

[Redmond 2007]  S. P. Redmond, "On zero-divisor graphs of small finite commutative rings", *Discrete Math.* **307**:9-10 (2007), 1155–1166.  MR  Zbl

[Vietri 2015]  A. Vietri, "A new zero-divisor graph contradicting Beck's conjecture, and the classification for a family of polynomial quotients", *Graphs Combin.* **31**:6 (2015), 2413–2423.  MR  Zbl

k.cooper@uky.edu                *Department of Mathematics, University of Kentucky, Lexington, KY, United States*

bpjohnson@fgcu.edu              *Department of Mathematics, Florida Gulf Coast University, Fort Myers, FL, United States*

msp

# The behavior of a population interaction-diffusion equation in its subcritical regime

Mitchell G. Davis, David J. Wollkind,
Richard A. Cangelosi and Bonni J. Kealy-Dichone

(Communicated by Martin J. Bohner)

A model interaction-diffusion equation for population density originally analyzed through terms of third-order in its supercritical parameter range is extended through terms of fifth-order to examine the behavior in its subcritical regime. It is shown that under the proper conditions the two subcritical cases behave in exactly the same manner as the two supercritical ones unlike the outcome for the truncated system. Further, there also exists a region of metastability allowing for the possibility of population outbreaks. These results are then used to offer an explanation for the occurrence of isolated vegetative patches and sparse homogeneous distributions in the relevant ecological parameter range where there is subcriticality for a plant-groundwater model system, as opposed to periodic patterns and dense homogeneous distributions occurring in its supercritical regime.

## 1. Introduction and formulation of the problem

Consider the following interaction-diffusion partial differential equation boundary value problem for $N = N(s, \tau) \equiv$ population density, where $s \equiv$ one-dimensional spatial variable and $\tau \equiv$ time:

$$\frac{\partial N}{\partial \tau} = D_0 \frac{\partial^2 N}{\partial s^2} + R_0 N_e r\left(\frac{N - N_e}{N_e}\right), \quad 0 < s < L, \tag{1-1a}$$

$$N(0, \tau) = N(L, \tau) = N_e, \tag{1-1b}$$

with

$$r(\theta) = \theta + \alpha\theta^3 + \gamma\theta^5 + O(\theta^7). \tag{1-1c}$$

Here, $D_0 \equiv$ dispersal constant, $R_0 \equiv$ interaction rate, $N_e \equiv$ equilibrium population density, and $L \equiv$ territory size, while $\alpha$ and $\gamma$ represent dimensionless interaction coefficients. Note that

$$N(s, \tau) \equiv N_e \tag{1-2}$$

is an exact solution to boundary value problem (1-1).

Introducing the nondimensional variables and parameter

$$z = \frac{\pi s}{L}, \quad t = \frac{D_0 \pi^2 \tau}{L^2}, \quad \theta(z, t) = \frac{N(s, \tau) - N_e}{N_e}, \quad \beta = \frac{R_0 L^2}{D_0 \pi^2}, \tag{1-3}$$

our original problem transforms into

$$\frac{\partial \theta}{\partial t} - \frac{\partial^2 \theta}{\partial z^2} = \beta r(\theta), \quad 0 < z < \pi, \tag{1-4a}$$

$$\theta(0, t) = \theta(\pi, t) = 0. \tag{1-4b}$$

Note that the exact solution (1-2) to the dimensional problem corresponds to

$$\theta(z, t) \equiv 0 \tag{1-5}$$

for our dimensionless one (1-4).

This is an extension to fifth-order of a model equation introduced by Wollkind et al. [1994] to illustrate the Stuart–Watson method of weakly nonlinear stability analysis of prototype reaction-diffusion equations. Asymptotic analyses of this sort are very useful for predicting pattern formation in such nonlinear systems. That analysis requires the expansion of $\theta$ in powers of an unknown function $A(t)$ with spatially dependent coefficients. The pattern-formational aspect of this system can be predicted from the long-time behavior of that amplitude function, which is governed by its Landau ordinary differential equation

$$\frac{dA}{dt} \sim \sigma A - a_1 A^3 - a_3 A^5 = F(A), \tag{1-6}$$

where $\sigma$ is the growth rate of linear stability theory and $a_{1,3}$ are the Landau constants. That long-time behavior is crucially dependent upon the signs of these Landau constants. Wollkind et al. [1994] concentrated on the special case for which $r(\theta) = \sin(\theta)$, employed by Matkowsky [1970] to develop his two-time method of weakly nonlinear stability theory, since their main concern was to compare the results obtained from the application of the Stuart–Watson method with those he deduced. Then $a_1 > 0$, identically (see below), and it is only necessary to include terms through third-order in $r(\theta)$ to make pattern formation predictions for this problem. In that event, there are two solutions of the truncated system: the first, a homogeneous one that is stable for $\sigma < 0$ and the second, a supercritical re-equilibrated pattern forming one that exists and is stable for $\sigma > 0$. These results

can be directly applied to our problem for its generalized $r(\theta)$ in the parameter range where $a_1 > 0$. In the range where $a_1 < 0$ and there is so-called subcriticality, the solutions to the truncated problem can grow without bound, and one must take the fifth-order terms into account in order to determine the long-time behavior of the system. Then we shall show that, if there is a parameter range over which the other Landau constant $a_3$ satisfies $a_3 > 0$, the pattern formation properties of our system can be ascertained without having to resort to considering even higher-order terms in $r(\theta)$. That requires the development of a formula for this Landau constant and an examination of its sign as a function of $\alpha$ and $\gamma$.

## 2. The Stuart–Watson method of nonlinear stability theory

Toward that end, we seek a Stuart–Watson expansion for the solution of our model equation of the form [Wollkind et al. 1994]

$$\theta(z, t) \sim A(t) \sin(z) + A^3(t)[\theta_{31} \sin(z) + \theta_{33} \sin(3z)]$$
$$+ A^5(t)[\theta_{51} \sin(z) + \theta_{53} \sin(3z) + \theta_{55} \sin(5z)]. \quad (2\text{-}1)$$

Note that the spatial terms in expansion (2-1) satisfy our boundary conditions (1-4b) at $z = 0$ and $\pi$, identically. Then, expanding $r(\theta)$ in powers of $A(t)$, employing the relevant trigonometric identities for the resulting products of sine functions contained in its coefficients, and making use of the Landau amplitude equation (1-6), we obtain a series of problems, one for each term appearing explicitly in our expansion of the form $A^n(t) \sin(mz)$, given by

$$A(t) \sin(z): \quad \sigma + 1 = \beta,$$
$$A^3(t) \sin(z): \quad 3\sigma\theta_{31} - a_1 + \theta_{31} = \beta\left(\theta_{31} + \tfrac{3}{4}\alpha\right),$$
$$A^3(t) \sin(3z): \quad 3\sigma\theta_{33} + 9\theta_{33} = \beta\left(\theta_{33} - \tfrac{1}{4}\alpha\right),$$
$$A^5(t) \sin(z): \quad 5\sigma\theta_{51} - a_3 - 3a_1\theta_{31} + \theta_{51} = \beta\left(\theta_{51} + \tfrac{9}{4}\alpha\theta_{31} - \tfrac{3}{4}\alpha\theta_{33} + \tfrac{5}{8}\gamma\right).$$

Although there are also two other $A^5(t)$ problems, they have not been cataloged above since only the one proportional to $\sin(z)$ which involves $a_3$ is required for our purposes. Here, while $\sigma$ and the $\theta_{nm}$ are being considered as functions of $\beta$, the coefficients $a_{1,3}$ are assumed to be independent of that bifurcation parameter and hence the use of the terminology Landau *constants*. That assumption is critical for their determination as solvability conditions, which is developed below.

We now solve these problems sequentially. Then, from the ones not involving these Landau constants, we obtain in a straightforward manner that

$$\sigma(\beta) = \beta - 1, \quad (2\text{-}2a)$$

and

$$\theta_{33}(\beta) = -\frac{\alpha\beta}{8(\beta+3)}, \tag{2-2b}$$

while the other two problems yield

$$2\sigma(\beta)\theta_{31}(\beta) = a_1 + \tfrac{3}{4}\alpha\beta \tag{2-2c}$$

and

$$4\sigma(\beta)\theta_{51}(\beta) = a_3 + 3\theta_{31}(\beta)\left(a_1 + \tfrac{3}{4}\alpha\beta\right) - \tfrac{3}{4}\alpha\beta\theta_{33}(\beta) + \tfrac{5}{8}\gamma\beta. \tag{2-2d}$$

(i) Assuming that $\theta_{31}(\beta)$ is well behaved at the critical bifurcation value of $\beta = 1$ and taking the limit of this first relation as $\beta \to 1$, while noting that $\sigma(\beta) = \beta - 1 \to 0$ in this limit, we obtain the solvability condition

$$a_1 = -\tfrac{3}{4}\alpha \tag{2-3a}$$

and, upon substitution of this back into (2-2c), the solution

$$\theta_{31}(\beta) \equiv \theta_{31} = \tfrac{3}{8}\alpha. \tag{2-3b}$$

Hence, we can deduce that

$$a_1 > 0 \quad \text{for } \alpha < 0 \qquad \text{and} \qquad a_1 < 0 \quad \text{for } \alpha > 0. \tag{2-4}$$

Thus, as mentioned earlier,

$$r(\theta) = \sin(\theta) = \theta - \tfrac{1}{6}\theta^3 + O(\theta^5) \quad \Longrightarrow \quad \alpha = -\tfrac{1}{6} \quad \Longrightarrow \quad a_1 = \tfrac{1}{8}. \tag{2-5}$$

Now, in this case, defining

$$\varepsilon^2 = \frac{\sigma(\beta)}{a_1} \quad \text{or} \quad \beta = 1 + \tfrac{1}{8}\varepsilon^2 \tag{2-6a}$$

and introducing the rescaled variables

$$\eta = \sigma t, \quad \mathcal{A}(\eta) = \frac{A(t)}{\varepsilon} \tag{2-6b}$$

into the truncated amplitude equation

$$\frac{dA}{dt} = \sigma A - a_1 A^3 + O(A^5), \tag{2-6c}$$

we obtain

$$\frac{d\mathcal{A}}{d\eta} = \mathcal{A} - \mathcal{A}^3 + O(\varepsilon^2), \tag{2-6d}$$

which justifies that truncation procedure. Now multiplying the truncated amplitude equation by $A(t)$ and rewriting it as

$$\frac{1}{2}\frac{dA^2}{dt} = \sigma A^2 - a_1 A^4 = \sigma A^2\left(1 - \frac{A^2}{\sigma/a_1}\right) = f_3(A^2), \tag{2-7}$$
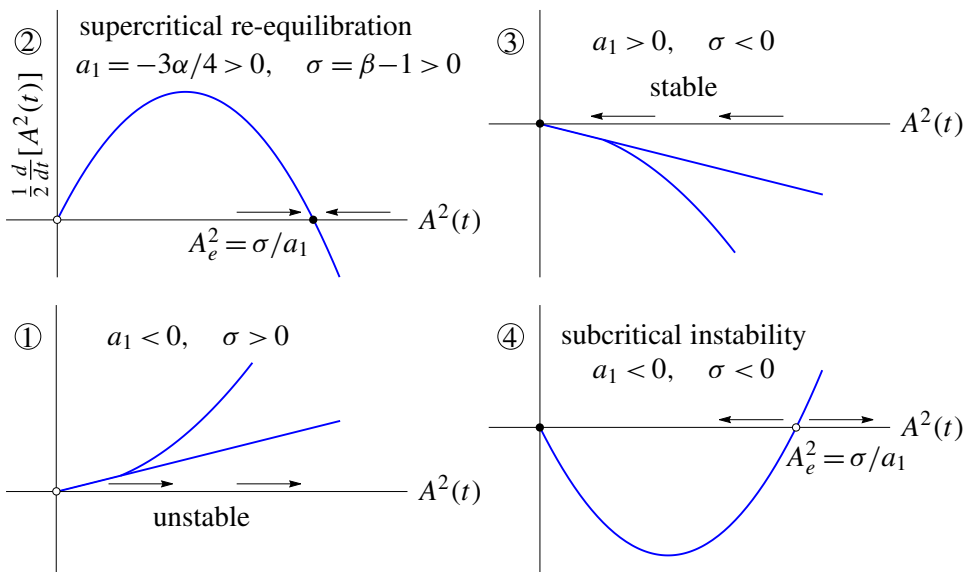
**Figure 1.** Plots of $f_3(A^2)$ for the third-order truncated amplitude equation with $\sigma = \beta - 1$ and $a_1 = -\frac{3}{4}\alpha$. Here the circled numbers correspond to the quadrants in the $\alpha\beta$-space of Figure 5 with horizontal axis $\beta = 1$ and vertical axis $\alpha = 0$.

we can easily deduce its long-time behavior by means of the four phase-plane plots of

$$\frac{1}{2}\frac{dA^2}{dt} = f_3(A^2)$$

that constitute Figure 1, which catalogs the four qualitatively different cases corresponding to the possibility of $\sigma$ and $a_1$ being either positive or negative. These serve as graphical representations of the cases discussed in Section 1 for the truncated version of our amplitude equation.

In particular, for the supercritical re-equilibration case of $\sigma, a_1 > 0$, we have

$$\lim_{t \to \infty} A(t) = A_e = \varepsilon, \tag{2-8a}$$

and hence

$$\lim_{t \to \infty} \theta(z, t) \sim \theta_e(z) = \delta \sin(z) \quad \text{as } \delta \to 0 \tag{2-8b}$$

since

$$\lim_{t \to \infty} \theta(z, t) = \varepsilon \sin(z) + \varepsilon^3[\theta_{31} \sin(z) + \theta_{33}(\beta) \sin(3z)] + O(\varepsilon^5)$$

$$= (\varepsilon + \theta_{31}\varepsilon^3) \sin(z) + \varepsilon^3 \theta_{33}(1) \sin(3z) + O(\varepsilon^5)$$

$$= \delta \sin(z) + \tfrac{1}{192}\delta^3 \sin(3z) + O(\delta^5) \sim \delta \sin(z) \quad \text{as } \delta \to 0, \tag{2-8c}$$
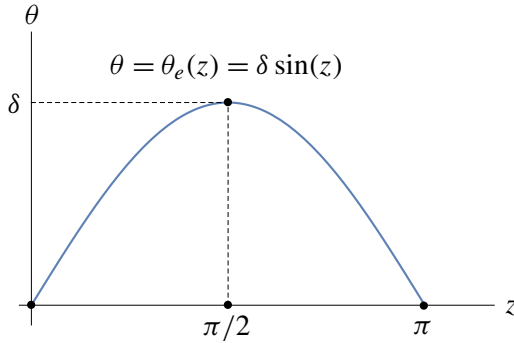
**Figure 2.** Plot of the arch solution $\theta_e(z)$ for $0 \le z \le \pi$.

where $\delta = \varepsilon + \varepsilon^3 \theta_{31} > 0$. This equilibrium state, plotted in Figure 2, is an arch-type pattern formed from one-cycle of a sine curve with its maximum amplitude $\delta$ occurring at $z = \frac{1}{2}\pi$.

(ii) We next proceed to analyze the second Landau constant relation (2-2d) involving $a_3$ and $\theta_{51}$ in an analogous manner to that just employed to evaluate $a_1$ and $\theta_{31}$. Thus, assuming $\theta_{51}(\beta)$ to be well behaved at $\beta = 1$ and taking the limit of this relation as $\beta \to 1$, we obtain the solvability condition

$$a_3 = -\tfrac{5}{8}\gamma - 3\theta_{31}\left(a_1 + \tfrac{3}{4}\alpha\right) + \tfrac{3}{4}\alpha\theta_{33}(1) = -\tfrac{5}{8}\gamma - \tfrac{3}{128}\alpha^2 \qquad (2\text{-}9a)$$

and, upon substitution of this back into (2-2d), the solution

$$\theta_{51}(\beta) = \tfrac{5}{32}\gamma + \tfrac{9}{16}\alpha\theta_{31} + \frac{3\alpha^2(4\beta + 3)}{512(\beta + 3)}. \qquad (2\text{-}9b)$$

Observe that, by virtue of the value of $a_1$, we have $a_3$ is independent of $\theta_{31}$. Also observe that, unlike this quantity, $\theta_{51}$ is a function of $\beta$. Finally note, in addition, should we have assumed that the Stuart–Watson expansion for $\theta(z, t)$ and the Landau equation for $dA/dt$ contained even powers of $A(t)$, then the solvability conditions and solutions for their coefficients would have shown them to be zero. Hence our implicit assumption that these quantities only contained odd powers was made without loss of generality and follows as a direct consequence of the form of $r(\theta)$.

Having determined its coefficients, we shall examine the truncated amplitude equation (1-6) through terms of fifth-order, i.e.,

$$\frac{dA}{dt} = F(A), \qquad (2\text{-}10)$$

and defer until after this examination has been completed a justification for that truncation. We seek conditions under which the inclusion of fifth-order terms will re-equilibrate the growing solutions predicted through third-order when $a_1 < 0$. Hence we assume a parameter range in which $a_1 < 0$ or $\alpha > 0$. Further, anticipating

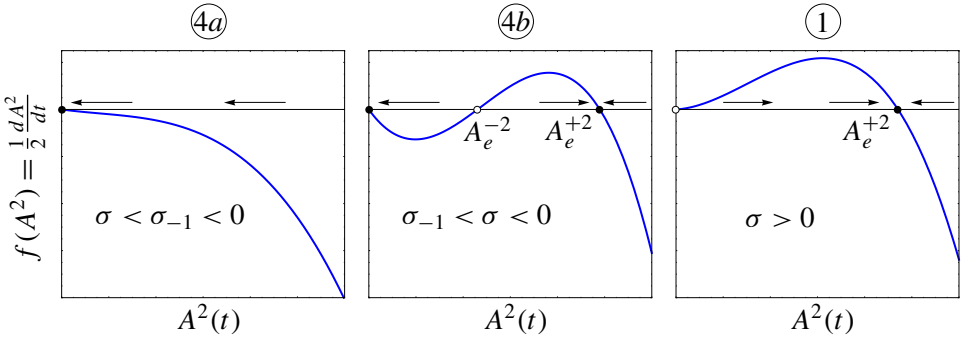**Figure 3.** Plots of $f(A^2)$ for the fifth-order truncated amplitude equation with $a_1 < 0$; $a_3 > 0$; and $\sigma < \sigma_{-1} = -a_1^2/(4a_3) < 0$, $\sigma_{-1} < \sigma < 0$, and $\sigma > 0$, respectively. Here, the circled numbers correspond to the quadrants in the $\alpha\beta$-space of Figure 5.

our results to be demonstrated below, we assume that $a_3 > 0$, while, as always, $\sigma \in \mathbb{R}$. This equation has three equilibrium points

$$A(t) \equiv A_e \quad \text{such that } F(A_e) = 0 \tag{2-11a}$$

satisfying either

$$A_e = 0 \quad \text{or} \quad 2a_3 A_e^{\pm 2} = \pm\sqrt{a_1^2 + 4a_3\sigma} - a_1. \tag{2-11b}$$

Observe that, since they must be real and positive, $A_e^{+2}$ exists for $\sigma \geq \sigma_{-1} = -a_1^2/(4a_3)$, while $A_e^{-2}$ only exists for $\sigma_{-1} \leq \sigma < 0$. Multiplying our truncated amplitude equation (2-10) by $A(t)$, we obtain

$$\frac{1}{2}\frac{dA^2}{dt} = \sigma A^2 - a_1 A^4 - a_3 A^6 = A^2(A_e^{-2} - A^2)(A^2 - A_e^{+2}) = f(A^2). \tag{2-12}$$

Then we can determine the global stability properties of these equilibrium points by plotting $\frac{1}{2}dA^2/dt = f(A^2)$ for $\sigma < \sigma_{-1} < 0$, $\sigma_{-1} < \sigma < 0$, and $\sigma > 0$, respectively, in the three phase-plane plots of Figure 3. From that figure, we can see that 0 is globally stable for $\sigma < \sigma_{-1} < 0$, $A_e^{+2}$ is globally stable for $\sigma > 0$, and in the overlap region where either can be stable, depending on initial conditions, 0 is stable for $0 < A^2(0) < A_e^{-2}$ and $A_e^{+2}$ is stable for $A^2(0) > A_e^{-2}$, while $A_e^{-2}$, which only exists in that bistability region, is not stable there.

To justify this truncation procedure we consider our Landau equation in the form

$$\frac{dA}{dt} = F(A) + O(A^7), \tag{2-13}$$

define $\varepsilon^2 = -a_1$, assume $a_3 = O(1)$ as $\varepsilon \to 0$, and let $\sigma = O(\varepsilon^4)$. Then $A_e^{+2} = O(\varepsilon^2)$, which implies that $A_e^+ = O(\varepsilon)$. Note, $\alpha = 10^{-2}$ and $\gamma = -2$ yield Landau constants
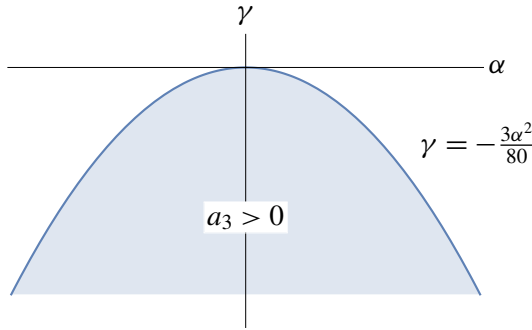
**Figure 4.** Plot of the region in the $\alpha\gamma$-plane, where $a_3 > 0$.

satisfying these conditions. Now, analogous to our approach at third order, we introduce the rescaled variables

$$\eta = \sigma t, \quad \mathcal{A}(\eta) = A(t)/A_e^+, \quad \text{where } \mathcal{A}, \frac{d\mathcal{A}}{d\eta} = O(1) \text{ as } \varepsilon \to 0. \qquad (2\text{-}14)$$

Since

$$\frac{dA}{dt} = \sigma A_e^+ \frac{d\mathcal{A}}{d\eta} = O(\varepsilon^5), \quad \sigma A = \sigma A_e^+ \mathcal{A} = O(\varepsilon^5),$$

$$a_1 A^3 = a_1 A_e^{+3} \mathcal{A}^3 = O(\varepsilon^5), \quad a_3 A^5 = a_3 A_e^{+5} \mathcal{A}^5 = O(\varepsilon^5), \qquad (2\text{-}15)$$

$$\text{while } O(A^7) = O(A_e^{+7} \mathcal{A}^7) = O(\varepsilon^7)$$

under these conditions, this justifies our truncation procedure at fifth order.

Finally, when $\sigma > 0$, we have the same type of equilibrium solution as depicted in Figure 2, except in this case

$$\delta = \varepsilon_0 + \theta_{31}(1)\varepsilon_0^3 + \theta_{51}(1)\varepsilon_0^5, \quad \text{where } A_e^+ = A_0\varepsilon = \varepsilon_0 \text{ with } A_0 = O(1) \text{ as } \varepsilon \to 0. \qquad (2\text{-}16)$$

This result depends upon

$$a_3 > 0 \implies \gamma < -\tfrac{3}{80}\alpha^2. \qquad (2\text{-}17)$$

Recall that, in addition, we have already taken $\alpha > 0$ to guarantee that $a_1 = -\tfrac{3}{4}\alpha < 0$. That region is plotted in the fourth quadrant of the $\alpha\gamma$-plane of Figure 4. In this context, note from Figure 3 that, unlike the situation depicted in Figure 1 for $\alpha > 0$, all the solutions remain bounded when the fifth-order terms in $r(\theta)$ are retained.

## 3. Bifurcation diagram, ecological interpretations, and conclusions

Should there exist a parameter range in a dynamical systems model of a given phenomenon for which the third-order Landau constant $a_1$ satisfies $a_1 < 0$ and hence the bifurcation is subcritical, the weakly nonlinear stability analysis must

be pushed to fifth order as originally pointed out by DiPrima et al. [1971]. This has been standard operating procedure particularly over the last five years when practitioners of the Palermo nonlinear stability theory group began considering fifth-order terms in the Landau equation during their investigation of subcritical bifurcation for a variety of two-component reaction-diffusion systems [Gambino et al. 2010; 2012; Tulumello et al. 2014]. By necessity, such calculations are long and technically complicated. Thus, when surveying the theory, there is some merit in introducing a simple model equation that preserves all the salient features of a more complex system but considerably reduces the labor involved in determining the Landau constants. This was our rationale for considering the generalized Matkowsky equation under investigation. That was also the rationale for Drazin and Reid's [1981] employment of their nondimensionalized version of the Matkowsky equation in order to develop weakly nonlinear theory relevant to hydrodynamic stability. Matkowsky [1970] regarded his problem as a mathematical model for temperature distribution in a finite bar with a nonlinear source term, the ends of which were maintained at the ambient, while Drazin and Reid [1981] offered their corresponding version as a phenomenological model of parallel flow in a channel. Hence, they both envisioned their instabilities to be rate-driven by considering the bifurcation parameter $\beta \sim R_0$. For ecological applications, it is often more relevant to envision these instabilities to be *territory-size* driven by considering $\beta \sim L^2$ and then the instability criterion describes the evolution of spatially heterogeneous structure in a specific domain.

Given that the fifth-order extensions referenced above primarily concentrated only on the subcritical regime, we begin this section by synthesizing our fifth-order results of Figure 3 valid for $a_1 < 0$ or, equivalently, $\alpha > 0$, and $a_3 > 0$ or, equivalently, $3\alpha^2 + 80\gamma < 0$, with those valid for $a_1 > 0$ or, equivalently, $\alpha < 0$, and $a_3 > 0$, as well. Note, that under these conditions, $A_e^{+2} > 0$ for $\sigma > 0$ and $A_e^{-2} < 0$, identically. If we plot a figure analogous to the supercritical cases of Figure 1, it is obvious that the qualitative morphological behavior of those cases is preserved at fifth order with the only change being now $A_e^2 = A_e^{+2}$. We accomplish this synthesis by means of Figure 5, a bifurcation diagram in $\alpha\beta$-space, where the relevant regions associated with these predicted morphological identifications are represented graphically. Since those results also depend on the behavior of $\sigma$, while $\sigma = 0$ and $\sigma = \sigma_{-1}$ are the critical loci for that quantity in this regard, it is necessary for us to generate loci equivalent to them in $\alpha\beta$-space. In this context, using our previous solvability conditions and definitions, we can deduce the following equivalences:

$$\sigma = \beta - 1 = 0 \quad \Longleftrightarrow \quad \beta = 1,$$

$$\sigma = \beta - 1 = \sigma_{-1} = -\frac{a_1^2}{4a_3} = \frac{18\alpha^2}{3\alpha^2 + 80\gamma} \quad \Longleftrightarrow \quad \beta = 1 + \frac{18\alpha^2}{3\alpha^2 + 80\gamma}, \tag{3-1}$$
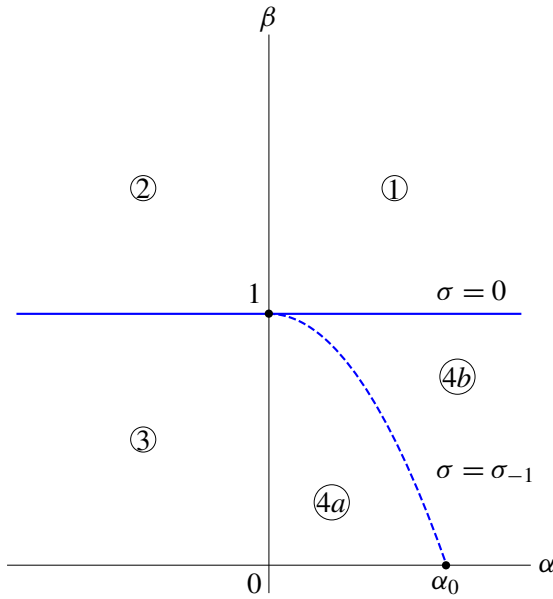
**Figure 5.** Bifurcation diagram in $\alpha\beta$-space with $\sigma_{-1} = -a_1^2/(4a_3)$, $\sigma = \beta - 1$, $a_1 = -\frac{3}{4}\alpha$, and $a_3 = -\frac{5}{8}\gamma - \frac{3}{128}\alpha^2 > 0$, where the circled numbers correspond to the quadrants denoted in Figures 1 and 3.

| quadrant | 1 | 2 | 3 | 4a | 4b |
|---|---|---|---|---|---|
| stable equilibrium point | $A_e^{+2}$ | $A_e^{+2}$ | 0 | 0 | $\begin{matrix}0\\A_e^{+2}\end{matrix}$ |

**Table 1.** Stable equilibrium points for $A^2$ in the quadrants of Figure 5.

which are plotted in Figure 5. Here, that first locus is a horizontal line parallel to the $\alpha$-axis which divides our $\alpha\beta$-space into the four quadrants formed by it and the $\beta$-axis, while the second is a concave downward decreasing curve having a horizontal tangent at its $\beta$-intercept of 1 and an $\alpha$-intercept of $\alpha_0 > 0$, where $\alpha_0^2 = -\frac{80}{21}\gamma$, which separates the fourth quadrant of that space into two parts. From an examination of the modification of the supercritical cases of Figure 1 described above and the subcritical cases of Figure 3, we construct Table 1 cataloging the stable equilibrium points for $A^2$ in each of the quadrants of Figure 5.

Note that these fifth-order results for our model equation are much more self-consistent than those obtained in the case of its third-order truncation, in that, the behavior for the subcritical quadrants 1 and 4a now exactly resemble the behavior for the supercritical quadrants 2 and 3, respectively. In the subcritical quadrant 4b, we have what biologists refer to as metastability, in that, the 0 equilibrium point is

| quadrant | 1 | 2 | 3 | 4a | 4b |
|---|---|---|---|---|---|
| stable pattern | arch | arch | dense hom. | sparse hom. | sparse hom. arch |

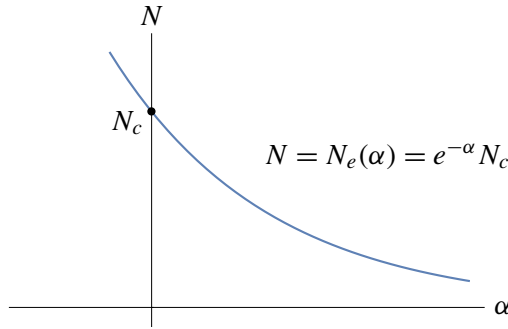**Table 2.** Morphological stability predictions for Table 1.



**Figure 6.** Plot of the population equilibrium density $N_e$ versus $\alpha$.

stable to initially small disturbances, but the model will switch to the equilibrium point $A_e^{+2}$ for sufficiently large ones. The existence of such a region of metastability allows our model equation to exhibit outbreak behavior wherein the maximum population level increases several-fold upon a sufficient initial perturbation in amplitude.

Returning to our original dimensional formulation (1-1), the fact that $A^2 = 0$ represents a globally stable equilibrium point implies that

$$\lim_{\tau \to \infty} N(s, \tau) = N_e. \tag{3-2}$$

Hence this solution represents a homogeneous population. In many actual biological systems, such as the interaction-diffusion plant-groundwater one employed by Chaiya et al. [2015] to model vegetative pattern formation in a flat arid environment, the homogeneous patterns in the subcritical parameter range correspond to relatively sparse distributions, while most of those patterns in the supercritical range correspond to much denser distributions, where the threshold between these two types of distributions occurs at some $N_c$. We can induce this sort of behavior in our model equation by adopting the relationship

$$N_e = N_e(\alpha) = N_c e^{-\alpha}, \tag{3-3}$$

which is plotted in Figure 6. Then from this relation and Table 1 in conjunction with Figure 2, we can deduce the stable pattern predictions given in Table 2 for the quadrants of Figure 5.

In [Chaiya et al. 2015], it was conjectured that the region of parameter space of subcriticality, where $a_1 < 0$, corresponded to isolated vegetative patches when $\sigma > 0$ and low-density homogeneous distributions when $\sigma < 0$, as opposed to the occurrence of periodic patterns for $\sigma > 0$ and high-density homogeneous distributions when $\sigma < 0$, where $a_1 > 0$, which were already predicted by their rhombic-planform two-dimensional nonlinear stability analysis. Such isolated patches are a compromise between periodic patterns and homogeneous stable states that are sparse enough to resemble bare ground. They then associated equilibrium points 0 and $A_e^{+2}$ of quadrants 1 and 4 of Table 1 with the sparse homogeneous state and the isolated patch, respectively, that would occur in a postulated fifth-order extension, should $a_3 > 0$ for this parameter range. Our fifth-order results summarized in Table 2 represent the first step in a conclusive demonstration of the validity of this conjecture.

We conclude by noting that although these results are only strictly asymptotically valid in a neighborhood of the marginal stability curve $\beta = 1$, Boonkorkuea et al. [2010], by comparing their theoretical predictions of this sort with existing numerical simulations of vegetative pattern formation for a model evolution equation, recently showed that the former can often be extrapolated to those regions of parameter space relatively far from the marginal curve. These theoretical predictions also associated that region of parameter space, where numerical simulation generated isolated patches, with $\sigma > 0$ and $a_1 < 0$.

Finally, we close by offering, for the sake of definiteness, a closed-form representation of $r(\theta)$, composed of combinations of common functions that produce Landau constants consistent in sign with our subcriticality assumptions. Recall the following Maclaurin polynomials truncated through terms of fifth order:

$$\sinh(z) \sim z + \tfrac{1}{6}z^3 + \tfrac{1}{120}z^5 \quad \text{and} \quad \arctan(z) \sim z - \tfrac{1}{3}z^3 + \tfrac{1}{5}z^5. \tag{3-4}$$

Then

$$\begin{aligned} 4\sinh\left(\tfrac{1}{2}\theta\right) &\sim 4\left(\tfrac{1}{2}\theta + \tfrac{1}{48}\theta^3 + \tfrac{1}{3840}\theta^5\right) = 2\theta + \tfrac{1}{12}\theta^3 + \tfrac{1}{960}\theta^5, \\ 2\arctan\left(\tfrac{1}{2}\theta\right) &\sim 2\left(\tfrac{1}{2}\theta - \tfrac{1}{24}\theta^3 + \tfrac{1}{160}\theta^5\right) = \theta - \tfrac{1}{12}\theta^3 + \tfrac{1}{80}\theta^5. \end{aligned} \tag{3-5}$$

Now, defining $r(\theta)$ to be the difference between these two functions, we obtain

$$\alpha = \tfrac{1}{6} > 0, \ \gamma = -\tfrac{11}{960} \quad \text{such that} \quad 80\gamma + 3\alpha^2 = -\tfrac{11}{12} + \tfrac{1}{12} = -\tfrac{5}{6} < 0. \tag{3-6}$$

## References

[Boonkorkuea et al. 2010] N. Boonkorkuea, Y. Lenbury, F. J. Alvarado, and D. J. Wollkind, "Nonlinear stability analyses of vegetative pattern formation in an arid environment", *J. Biol. Dyn.* **4**:4 (2010), 346–380. MR Zbl

[Chaiya et al. 2015] I. Chaiya, D. J. Wollkind, R. A. Cangelosi, B. J. Kealy-Dichone, and C. Rattanakul, "Vegetative rhombic pattern formation driven by root suction for an interaction-diffusion

plant-ground water model system in an arid flat environment", *Amer. J. Plant Sci.* **6**:8 (2015), 1278–1300.

[DiPrima et al. 1971] R. C. DiPrima, W. Eckhaus, and L. A. Segel, "Non-linear wave-number interaction in near-critical two-dimensional flows", *J. Fluid Mech.* **49**:4 (1971), 705–744. Zbl

[Drazin and Reid 1981] P. G. Drazin and W. H. Reid, *Hydrodynamic stability*, Cambridge Univ. Press, 1981. MR Zbl

[Gambino et al. 2010] G. Gambino, A. M. Greco, M. C. Lombardo, and M. Sammartino, "A subcritical bifurcation for a nonlinear reaction-diffusion system", pp. 163–172 in *Waves and stability in continuous media* (Palermo, 2009), edited by A. M. Greco et al., World Sci. Publ., Hackensack, NJ, 2010. MR Zbl

[Gambino et al. 2012] G. Gambino, M. C. Lombardo, and M. Sammartino, "Turing instability and traveling fronts for a nonlinear reaction-diffusion system with cross-diffusion", *Math. Comput. Simulation* **82**:6 (2012), 1112–1132. MR Zbl

[Matkowsky 1970] B. J. Matkowsky, "A simple nonlinear dynamic stability problem", *Bull. Amer. Math. Soc.* **76** (1970), 620–625. MR Zbl

[Tulumello et al. 2014] E. Tulumello, M. C. Lombardo, and M. Sammartino, "Cross-diffusion driven instability in a predator-prey system with cross-diffusion", *Acta Appl. Math.* **132** (2014), 621–633. MR Zbl

[Wollkind et al. 1994] D. J. Wollkind, V. S. Manoranjan, and L. Zhang, "Weakly nonlinear stability analyses of prototype reaction-diffusion model equations", *SIAM Rev.* **36**:2 (1994), 176–214. MR Zbl

m.g.davis@outlook.com          *Department of Mathematics, Washington State University, Pullman, WA, United States*

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

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

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

# Forbidden subgraphs of coloring graphs

Francisco Alvarado, Ashley Butts,
Lauren Farquhar and Heather M. Russell

(Communicated by Jerrold Griggs)

Given a graph $G$, its $k$-coloring graph has vertex set given by the proper $k$-colorings of the vertices of $G$ with two $k$-colorings adjacent if and only if they differ at exactly one vertex. Beier et al. (*Discrete Math.* **339**:8 (2016), 2100–2112) give various characterizations of coloring graphs, including finding graphs which never arise as induced subgraphs of coloring graphs. These are called forbidden subgraphs, and if no proper subgraph of a forbidden subgraph is forbidden, it is called minimal forbidden. In this paper, we construct a finite collection of minimal forbidden subgraphs that come from modifying theta graphs. We also construct an infinite family of minimal forbidden subgraphs similar to the infinite family found by Beier et al.

## 1. Introduction

A *graph* $G = (V, E)$ consists of a set $V = V[G] = \{v_1, \dots, v_n\}$ of *vertices* and a set $E = E[G] \subseteq \{vv' : v, v' \in V\}$ of *edges*, where $vv'$ represents an unordered pair of vertices. In this paper, we assume $G$ has finite order (i.e., $|V|$ is finite), $v \neq v'$ whenever $vv' \in E$, and $G$ has at most one edge between a single pair of vertices. A graph $H$ is an *induced subgraph of $G$* if $V[H] \subseteq V[G]$ and $vv' \in E[H]$ if and only if $vv' \in E[G]$.

Given $k \in \mathbb{N}$, a *proper $k$-coloring* of a graph $G$ is a function $\alpha : V[G] \to \{1, 2, \dots, k\}$ such that $\alpha(v) \neq \alpha(v')$ whenever $vv' \in E[G]$. The *$k$-coloring graph of $G$*, denoted by $\mathcal{C}_k(G)$, is the graph with vertex set consisting of all proper $k$-colorings of $G$. Edges between colorings exist if and only if the colorings differ at precisely one vertex of $G$. Figure 1 shows an example. When discussing properties of $\mathcal{C}_k(G)$, we refer to $G$ as the base graph for $\mathcal{C}_k(G)$.

Interest in coloring graphs stems from applications in theoretical physics. Coloring graphs model the Glauber dynamics of the antiferromagnetic Potts model at zero temperature [Dyer et al. 2006; Jerrum 1995; Molloy 2004; Vigoda 2000]. Beier et al. [2016] approach coloring graphs from an inverse perspective, asking "Given a

graph $G'$, does there exist a graph $G$ and natural number $k$ such that $\mathcal{C}_k(G) = G'$?"
We build on their work on permissible and forbidden subgraphs of coloring graphs.

A graph $H'$ is called *permissible* if it is an induced subgraph of some coloring graph. If $H'$ is not an induced subgraph of any coloring graph, we say $H'$ is *forbidden*. The graph $H'$ is called *minimal forbidden* if $H'$ is forbidden and each proper induced subgraph of $H'$ is permissible. Beier et al. define an infinite two-parameter family of graphs $M_{n,p}$ and show an infinite number of them are minimal forbidden. They define another infinite collection of graphs called theta graphs and completely classify them into permissible, minimal forbidden, and forbidden but not minimal.

The goal of this paper is to formalize and enhance the tools and techniques for studying the forbidden and permissible subgraphs of coloring graphs introduced in [Beier et al. 2016] and to provide new examples. This will aid others investigating coloring graphs and, perhaps more interestingly, other types of transition graphs, like those found in [Cohen and Teicher 2014; Zhang et al. 1988; Haas 2012; Mohar 2007].

Section 2 expands on coloring edge labeling and edge labeling partitions, which were first introduced in [Beier et al. 2016]. We also recall necessary results from that paper involving permissible subgraphs. As an application of Section 2, we give two new collections of minimal forbidden subgraphs in Section 3. One collection comes from modifying theta graphs, and the other is an infinite subset of the two-parameter family of graphs $L_{n,p}$, which we define in that section. Finally, Section 4 provides several future directions for this work.

Our notation and conventions follow [Beier et al. 2016; Diestel 1997]. If we are unsure whether a graph is a coloring graph, we sometimes refer to it as a *candidate coloring graph*. Base graphs will be denoted by $G$, and candidate coloring graphs will be denoted by $G'$. Subgraphs (usually induced) of $G$ and $G'$ will be denoted by $H$ and $H'$ respectively. Vertices in the coloring graph will be identified by Greek letters ($\alpha, \beta, \gamma, \dots$), and vertices in the base graph will be denoted by lowercase letters ($u, v, w, \dots$).

We denote by $I_n$ the graph consisting of $n$ vertices and no edges. Given graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, we denote their disjoint union by $G_1 \sqcup G_2$. The Cartesian product of $G_1$ and $G_2$, denoted by $G_1 \square G_2$, has vertex set $V_1 \times V_2$ with an edge between $(v_1, v_2)$ and $(w_1, w_2)$ exactly when $v_1 = w_1$ and $v_2 w_2 \in E_2$ or $v_2 = w_2$ and $v_1 w_1 \in E_1$.

## 2. Background

In this section, we recall and formalize definitions, theorems, and techniques from [Beier et al. 2016] needed to analyze forbidden and permissible subgraphs. We begin with a discussion of edge labeling of coloring graphs, which is a key tool used to prove graphs are minimal forbidden.
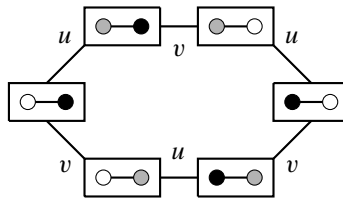
**Figure 1.** Coloring edge labeling of $\mathcal{C}_3(P_1)$.

An *edge labeling* of a graph is a function with domain the edge set of the graph. Given a graph $G$ and $k \in \mathbb{N}$, the *coloring edge labeling* of the coloring graph $\mathcal{C}_k(G)$ is the map from $E[\mathcal{C}_k(G)]$ to $V[G]$ that labels each edge $\alpha\beta \in E[\mathcal{C}_k(G)]$ with the unique vertex of $V[G]$ at which the colorings $\alpha$ and $\beta$ differ. This labeling technique was first introduced in [Beier et al. 2016, p. 2102], where it is referred to as edge labeling. Figure 1 shows the coloring edge labeling for coloring graph $\mathcal{C}_3(P_1)$, where $V[P_1] = \{u, v\}$.

For a graph $H'$, we call an edge labeling a *proper edge labeling* if there exists a graph $G$ and a $k \in \mathbb{N}$ such that $H'$ is an induced subgraph of $\mathcal{C}_k(G)$ and the edge labeling of $H'$ coincides with the coloring edge labeling of $\mathcal{C}_k(G)$. An *improper edge labeling* is an edge labeling that is not proper.

It follows from these definitions that a graph $H'$ is permissible if and only if it has a proper edge labeling. In [Beier et al. 2016, Corollary 12], it is shown that all cycles except $C_5$ are permissible subgraphs, so a cycle $C_n$ of size $n \neq 5$ must have at least one proper edge labeling. We use properties of proper edge labelings of cycles to analyze proper edge labelings of more complicated graphs. The following lemma summarizes properties of coloring edge labelings of cycles used in [Beier et al. 2016].

**Lemma 1.** *A proper edge labeling of a cycle $C_n$ must satisfy the following conditions*:

(1) *Each label must occur at least twice.*

(2) *Adjacent edges have the same label if and only if $n = 3$.*

(3) *If a cycle has three edges consecutively labeled $u, v, u$ with $u \neq v$ then either $n = 4$ or $u$ occurs as a label at least three times.*

While the conditions outlined in Lemma 1 are necessary for a proper edge labeling of a cycle, they are not sufficient. One can show that the edge labeling in Figure 2 is not proper though it meets all conditions in Lemma 1. Also, we emphasize that the conditions in Lemma 1 are necessary only for cycles.

Also introduced in [Beier et al. 2016, p. 2102] is the concept of edge label partitioning. The *edge label partition* corresponding to a proper edge labeling of a cycle $C_n$ is the partition of $n$ consisting of the number of occurrences of each label.
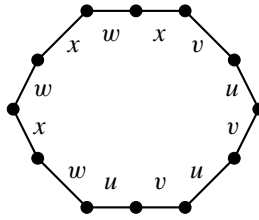
**Figure 2.** An improper edge labeling of $C_{12}$.

| $C_3$ | $C_4$ | $C_6$ | $C_7$ | $C_8$ | $C_9$ |
|---|---|---|---|---|---|
| $3 \vdash 3$ | $4 \vdash 2, 2$ | $6 \vdash 3, 3$ | $7 \vdash 2, 2, 3$ | $8 \vdash 4, 4$ | $9 \vdash 2, 3, 4$ |
| | | $6 \vdash 2, 2, 2$ | | $8 \vdash 2, 2, 4$ | $9 \vdash 3, 3, 3$ |
| | | | | $8 \vdash 2, 3, 3$ | $9 \vdash 2, 2, 2, 3$ |
| | | | | $8 \vdash 2, 2, 2, 2$ | |

**Table 1.** Edge label partition types for small cycles.

For example, the edge label partition corresponding to the proper edge labeling of $C_6$ shown in Figure 1 is $6 \vdash 3, 3$ since $u$ and $v$ each occur three times. Note that each proper edge labeling corresponds to a unique edge label partition. However, a partition does not necessarily uniquely determine a proper edge labeling.

Moreover, not every partition of $n$ corresponds to a proper edge labeling. In fact, conditions on edge labelings stated in Lemma 1 give restrictions on which partitions can be edge label partitions. Each part of an edge label partition must be greater than 1 according to the first condition in Lemma 1. Also by the first condition, no part of an edge label partition can be greater than half of $n$. The following is a complete list of possible edge label partition types for $C_n$ with $3 \le n \le 9$. (Recall $C_5$ is forbidden, so there are no edge label partitions of 5.)

Table 1 is very useful when attempting to find proper edge labelings of graphs. For instance, if $H'$ contains an induced copy of $C_7$, then a proper edge labeling of $H'$ must have exactly three distinct labels on that cycle and a corresponding edge label partition type of $7 \vdash 2, 2, 3$. We can then use Lemma 1 to further investigate how those labels could be arranged.

In addition to examining cycles, our analysis of forbidden subgraphs builds on the following results about permissible and forbidden subgraphs.

**Theorem 2** [Beier et al. 2016, Theorem 9]. *If $H_1'$ and $H_2'$ are permissible, then $H_1' \sqcup H_2'$ is permissible. Alternately, if $H_1' \sqcup H_2'$ is forbidden, then either $H_1'$ or $H_2'$ is forbidden.*

**Theorem 3** [Beier et al. 2016, Theorem 10]. *If $H_1'$ and $H_2'$ are permissible, then $H_1' \square H_2'$ is permissible. Alternately, if $H_1' \square H_2'$ is forbidden, then either $H_1'$ or $H_2'$ is forbidden.*
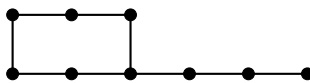
**Figure 3.** Attaching $P_3$ to $C_6$.

These two preceding theorems allow us to construct a number of permissible subgraphs. Since the path $P_1$ is permissible, given any permissible subgraph $H'$, it follows that $P_1 \square H'$ is permissible. Note that $H'$ with one new vertex and an edge from that new vertex to any other vertex is an induced subgraph of $P_1 \square H'$. We call this *attaching a copy of $P_1$*. More generally, we refer to the process of identifying an endpoint of a path with some vertex of a graph $H'$ as *attaching a path to $H'$*. Figure 3 shows an example of attaching a path to a 6-cycle. By an inductive argument, we arrive at Corollary 4. By similar arguments, Corollaries 5 and 6 also follow from Theorems 2 and 3.

**Corollary 4** [Beier et al. 2016, p. 2104]. *A permissible subgraph with any number of paths of any length attached is permissible.*

**Corollary 5** [Beier et al. 2016, Corollary 11]. *All trees are permissible.*

**Corollary 6** [Beier et al. 2016, Corollary 12]. *The graph $C_n$ for $n \neq 5$ is permissible. The graph $C_5$ is forbidden.*

In addition to appending paths to build new permissible subgraphs, we can sometimes add additional vertices along induced paths of permissible subgraphs to get new permissible subgraphs. The next theorem explains the conditions under which this can be done. The result of replacing an edge of a graph with $P_2$ will be called *subdividing an edge*.

**Theorem 7** [Beier et al. 2016]. *Let $H'$ be a permissible subgraph containing a degree-2 vertex whose neighbors are not adjacent. The graph obtained by subdividing both edges incident to the vertex of degree 2 is also permissible.*

This subdivision theorem is useful when studying permissibility of so-called theta graphs. A (*generalized*) *theta graph*, denoted by $T(m_1, m_2, \ldots, m_k)$ where $m_i \leq m_{i+1}$ for all $i$, consists of a collection of internally disjoint paths of lengths $m_1$, $m_2, \ldots, m_k$ with a single common initial vertex $u$ and terminal vertex $v$ where $u \neq v$. Thus, $u$ and $v$ will have degree $k$, while all other vertices have degree 2. Note that theta graphs generalize cycles since $C_n = T(1, n-1)$. The collection of generalized theta graphs are completely categorized as permissible, minimal forbidden, or forbidden not minimal in [Beier et al. 2016]. Any theta graph not containing those listed in the following theorem is permissible.

**Theorem 8** [Beier et al. 2016, Theorem 15]. *The complete list of minimal forbidden theta graphs is*

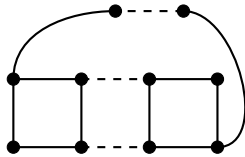$$T(1, 4), \quad T(1, 2, 2), \quad T(2, 2, 2), \quad T(3, 3, 3) \quad \text{and} \quad T(2, 2, 4).$$

**Figure 4.** $M_{n,p}$, where $n$, $p \geq 1$.

An infinite set of minimal forbidden subgraphs is introduced in [Beier et al. 2016]. These are part of the set of graphs denoted by $M_{n,p}$, where $n$, $p \geq 1$. These graphs contain a chain of $n-1$ induced copies of $C_4$ with a path of length $p+1$ between two vertices, as seen in Figure 4. The following theorem, which is needed in our arguments, summarizes the results on $M_{n,p}$ graphs from [Beier et al. 2016].

**Theorem 9** [Beier et al. 2016, Lemma 16, Theorem 17]. *The family $M_{n,p}$ is forbidden but not minimal if and only if $n \geq 1$ and $p \leq 2$. The family $M_{n,3}$ is minimal forbidden if and only if $n \geq 2$.*

## 3. Two collections of minimal forbidden subgraphs

Figure 5 has 14 new examples of minimal forbidden subgraphs that come from modifying the structure of generalized theta graphs. We will prove graph (a) is minimal forbidden by applying the language and lemmas from the previous section. The proofs that the other graphs are minimal forbidden are very similar in style and are therefore left to the reader.

The proof that a graph is minimal forbidden breaks into two parts: showing it is forbidden and showing it is minimal. To prove a graph is forbidden, we focus on its induced cycles examining their interactions and showing they have no simultaneous proper edge labelings. To prove a forbidden subgraph is minimal, we show that each of its proper induced subgraphs is permissible. Since subgraphs of permissible subgraphs are permissible, it is sufficient to show that all induced subgraphs obtained by removing one vertex are permissible.

**Theorem 10.** *Graph (a) in Figure 5 is a minimal forbidden subgraph.*

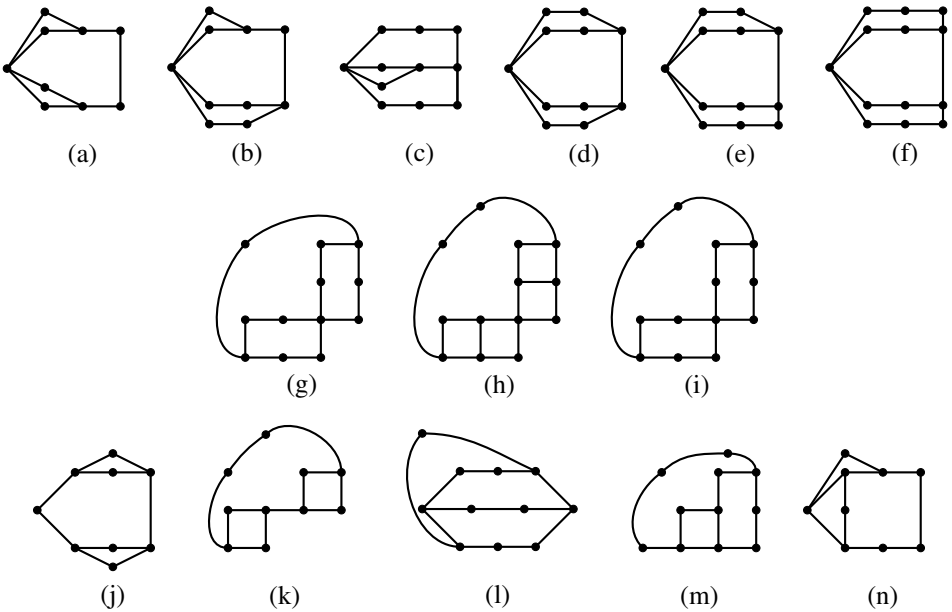*Proof.* Consider the following graph $H'$, which is Figure 5(a) with a choice of vertex names:
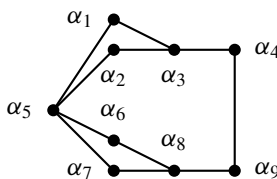
**Figure 5.** A finite collection of minimal forbidden subgraphs.

Since $\alpha_5$ is not a vertex of an induced $C_3$, a proper edge labeling of $H'$ must assign a different label to each of the four edges incident to $\alpha_5$. Since the edges incident to $\alpha_5$ are part of two edge-disjoint induced copies of $C_4$, these induced copies of $C_4$ will have repeated edge labels, say $u, v, u, v$ and $w, x, w, x$. However, each induced copy of $C_4$ shares two consecutive edges with an induced copy of $C_7$. Thus this $C_7$ would have at least four distinct edge labels, which is not possible by Lemma 1. We conclude $H'$ is forbidden.

Next, we demonstrate that $H'$ is permissible by arguing that removing any vertex from $H'$ yields a permissible subgraph. Removing $\alpha_1$, $\alpha_2$, $\alpha_6$, or $\alpha_7$ from $H'$ results in a copy of theta graph $T(2, 2, 5)$, which is permissible by Theorem 8. Removing $\alpha_5$ forms a tree, which is permissible by Corollary 5. Upon removing $\alpha_3$ or $\alpha_8$ from $H'$ we obtain $C_4$ with three paths of length 1 or 2 attached. Such graphs are permissible by Theorem 3 and Corollary 5. Removing $\alpha_4$ or $\alpha_9$ results in a tree attached to two copies of $C_4$ that share a vertex. Two copies of $C_4$ glued at one vertex is an induced subgraph of $C_4 \square P_2$, which is permissible by Theorem 3. Appending a tree is then permissible by Corollary 5. It follows that all proper induced subgraphs of $H'$ are permissible, and so $H'$ is a minimal forbidden subgraph. $\square$

We now construct a new infinite family of minimal forbidden subgraphs similar to the subset of $M_{n,p}$ graphs discussed in [Beier et al. 2016]. Figure 6 shows the graph $L_{n,p}$ with $n, p \geq 0$; the vertex names in the figure will be referenced in our arguments.

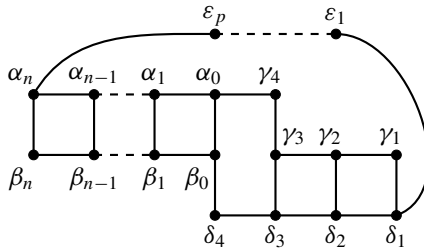**Figure 6.** Vertex labels for $L_{n,p}$, where $n, p \geq 0$.

**Lemma 11.** *For all $p \geq 1$ and $n \geq 0$ with $p + n \geq 3$, the graph $L_{n,p}$ is permissible.*

*Proof.* We begin by arguing that $L_{2k,1}$ is an induced subgraph of $\mathcal{C}_{k+1}(I_4)$ and $L_{2k+1,1}$ is an induced subgraph of $\mathcal{C}_{k+2}(I_4)$ for $k \geq 1$. Note that we include 0 as a color. Consider $I_4$ with $V[I_4] = \{u, v, w, y\}$. We represent colorings of $I_4$ as sequences of four numbers. For instance, the sequence 1230 corresponds to the coloring $\alpha$ where $\alpha(u) = 1$, $\alpha(v) = 2$, $\alpha(w) = 3$, and $\alpha(y) = 0$. Figure 7 shows a set of colorings of $I_4$ using $k + 1$ colors that span a copy of $L_{2k,1}$ in $\mathcal{C}_{k+1}(I_4)$.

For $n = 2k + 1$, the construction is almost the same. Consider the colorings in Figure 7 with the following modifications. Add colorings $00k(k+1)$ and $10k(k+1)$ on the left. Change the top coloring from $01kk$ to $01k(k+1)$ and the rightmost colorings from $22kk$ and $21kk$ to $22k(k+1)$ and $21k(k+1)$. One can check that these colorings span a copy of $L_{2k+1,1}$ in $\mathcal{C}_{k+2}(I_4)$.

Consider $L_{n,p}$ with $p > 1$, $n \geq 0$, and $p + n \geq 3$. Then $n + p - 1 \geq 2$, and hence it follows by the previous argument that $L_{n+p-1,1}$ is permissible. Removing all vertices $\beta_i$ from $L_{n+p-1,1}$ with $n+1 \leq i \leq n+p-1$ yields an induced copy of $L_{n,p}$. Induced subgraphs of permissible subgraphs are permissible, so $L_{n,p}$ is permissible. $\square$

For $n + p < 3$, the graphs $L_{n,p}$ are forbidden but not minimal. Indeed, the graph $L_{0,0}$ contains an induced copy of $T(2, 2, 4)$, the graph $L_{0,1}$ contains an induced
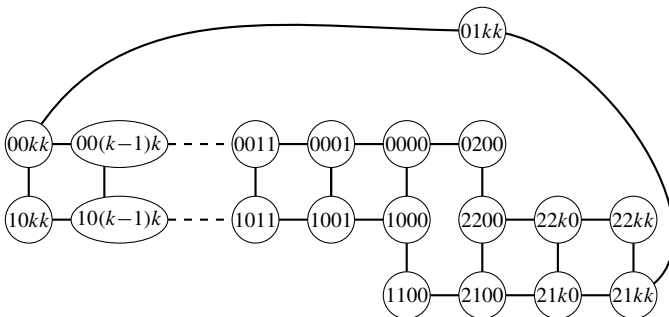


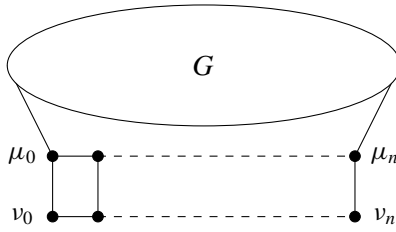**Figure 7.** $L_{2k,1}$ is an induced subgraph of $\mathcal{C}_{k+1}(I_4)$.

**Figure 8.** The graph $G'$ described in Lemma 12.

copy of $M_{2,3}$, and the graph $L_{1,0}$ contains an induced copy of $M_{3,3}$. The graphs $L_{2,0}$, $L_{0,2}$, and $L_{1,1}$ each contain an induced copy of graph (d) in Figure 5. Our final goal is to show that $L_{n,0}$ is minimal forbidden for all $n \geq 3$, but first we need one additional lemma.

**Lemma 12.** *Let G be a permissible subgraph containing an induced path of length n on vertices $\mu_0, \ldots, \mu_n$. Then the graph $G'$ with $V[G'] = V[G] \cup \{v_0, \ldots, v_n\}$ and*

$$E[G'] = E[G] \cup \{v_i v_{i+1} : 0 \leq i < n\} \cup \{\mu_i v_i : 0 \leq i \leq n\}$$

*is also permissible. The graph $G'$ is shown in Figure 8.*

*Proof.* Since $G$ is permissible, the graph $G \square P_1$ is permissible. The graph $G'$ is an induced subgraph of $G \square P_1$.                                                                                 □

**Theorem 13.** *For $n \geq 3$, the graph $L_{n,0}$ is minimal forbidden.*

*Proof.* Any proper edge labeling of $L_{n,0}$ must restrict to a proper edge labeling of the central copy of $C_6$ spanned by $\alpha_0$, $\beta_0$, $\gamma_3$, $\gamma_4$, $\delta_3$ and $\delta_4$. By Lemma 1, the only possible proper edge labelings of $C_6$ are (a) $u, v, u, v, u, v$ or (b) $u, w, v, u, w, v$, where $u$, $v$, and $w$ are distinct vertices in a base graph. Without loss of generality, assume edge $\beta_0\alpha_0$ has label $u$. This is illustrated in Figure 9.

In case (a), edges $\alpha_0\gamma_4$, $\delta_4\beta_0$, and $\gamma_3\delta_3$ have label $v$, while edges $\gamma_3\gamma_4$ and $\delta_3\delta_4$ have label $u$. By Lemma 1, all edges $\alpha_i\beta_i$ for $1 \leq i \leq n$ must have label $u$, and edges $\delta_j\gamma_j$ for $1 \leq j \leq 3$ must have label $v$. Furthermore $u$ must label at least one more edge in the induced $(n+6)$-cycle highlighted in Figure 9.

Invoking Lemma 1 once again, we see that $u$ cannot label $\alpha_{i-1}\alpha_i$ for any $1 \leq i \leq n$ or edges $\alpha_n\delta_1$ and $\gamma_2\gamma_3$ since $u$ labels an adjacent edge in each case. Thus $u$ can only label edge $\gamma_1\gamma_2$. This contradicts the third statement in Lemma 1 since the cycle under consideration has size greater than 4. We conclude that a proper edge labeling of $L_{n,0}$ restricting to the labeling of $C_6$ in case (a) does not exist.

In case (b) by Lemma 1, the edges $\alpha_i\beta_i$ for $1 \leq i \leq n$ and $\gamma_j\delta_j$ for $1 \leq j \leq 3$ must have label $u$. With these forced edge labelings, the $(n+6)$-cycle shown in Figure 9 does not have a proper edge labeling satisfying the conditions of Lemma 1
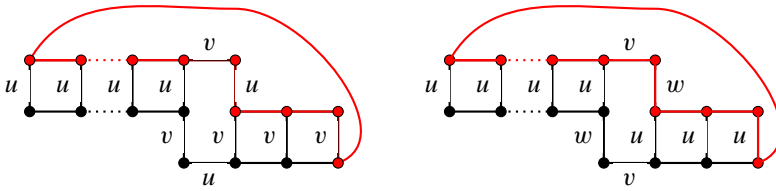
**Figure 9.** Possible edge labelings for $L_{n,0}$ from the proof of Theorem 13, case (a) on the left and case (b) on the right.

since $u$ cannot label two edges. We conclude that a proper edge labeling of $L_{n,0}$ restricting to the labeling of $C_6$ in case (b) does not exist. Since no proper edge labeling of $L_{n,0}$ restricts to a proper edge labeling of the central copy of $C_6$, we conclude that $L_{n,0}$ has no proper edge labeling and is therefore forbidden.

We now check that each induced subgraph of $L_{n,0}$ spanned by all but one vertex is permissible. We refer to the vertex labels of $L_{n,0}$ shown in Figure 6. There are seven cases.

Case 1: Removing $\delta_1$ or $\alpha_n$ yields a 6-cycle with two disjoint chains of 4-cycles, one of which has an attached copy of $P_1$. Recall that 6-cycles are permissible by Corollary 6 and attaching paths to permissible subgraphs yields new permissible subgraphs by Corollary 4. Finally, note that by inductively applying Lemma 12 to a path of length 1, one can append a chain of 4-cycles to a permissible subgraph to obtain another permissible subgraph.

Case 2: Removing $\beta_0$, $\alpha_0$, $\delta_3$, $\delta_4$, $\gamma_3$, or $\gamma_4$ yields a proper induced subgraph of $M_{n+4,0}$ possibly with paths attached. The graph $M_{n+4,0}$ is permissible by Theorem 9, so every induced subgraph is also permissible. Once again, Corollary 4 allows us to attach a paths.

Case 3: Removing $\beta_i$ for $1 \leq i \leq n$ yields a copy of $P_1$ attached to $L_{i-1,n-i+1}$ with a chain of 4-cycles attached along a path of length $n - i$, as in Lemma 12. Since $n \geq 3$ and $i \geq 1$, it follows that $(i - 1) + (n - i + 1) \geq 3$. Thus by Section 3 we see that $L_{i-1,n-i+1}$ is permissible.

The remaining cases are proven with explicit constructions. In the figures, strings of lengths 3 and 4 represent colorings of $I_3$ and $I_4$. If $n$ is even, we say $n = 2k$, and if $n$ is odd, say $n = 2k + 1$. Since $n \geq 3$, we have $k \geq 1$ for $n$ even or odd.

Case 4: For $n \geq 3$, the graph spanned by all but vertex $\delta_2$ of $L_{n,0}$ is an induced subgraph of $C_{k+3}(I_3)$, as is shown in Figure 10 for $n$ even and Figure 11 for $n$ odd.

Case 5: If $n$ is odd, the subgraph of $L_{n,0}$ spanned by all but vertex $\gamma_1$ is an induced subgraph of $C_{k+4}(I_3)$, as is shown in Figure 12. For $n$ even, the subgraph of $L_{n,0}$ spanned by all but vertex $\gamma_1$ is an induced subgraph of $C_{k+2}(I_4)$, as is shown in Figure 13.
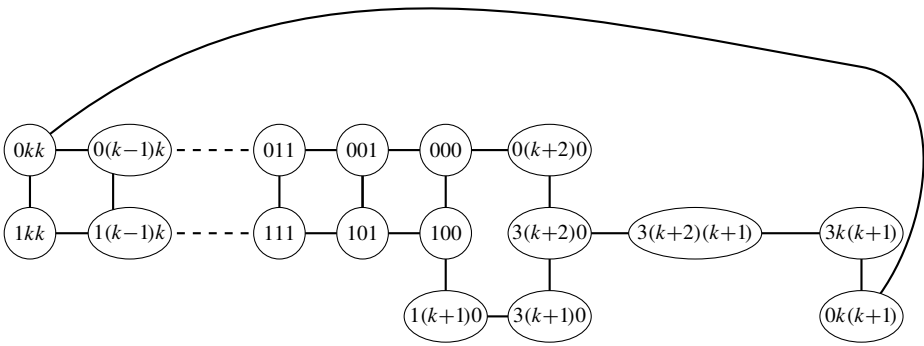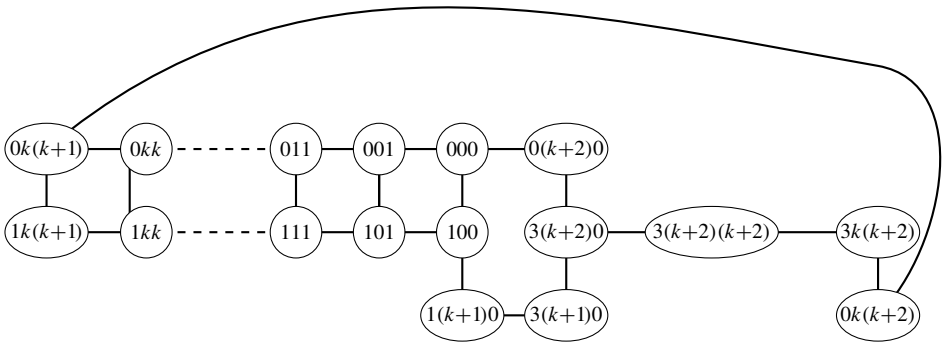
**Figure 10.** Case 4 for *n* even.
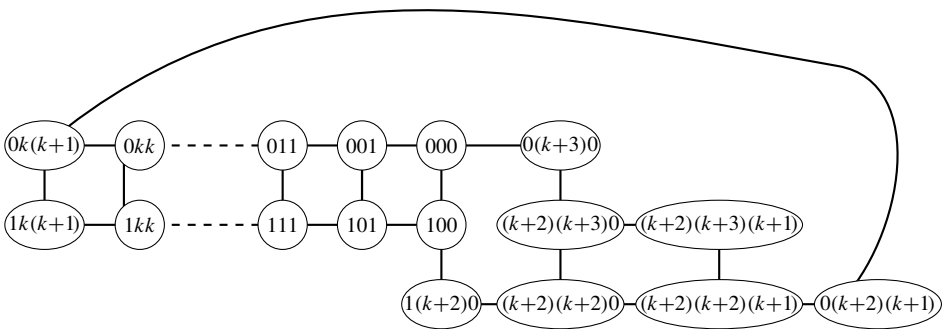


**Figure 11.** Case 4 for *n* odd.



**Figure 12.** Case 5 for *n* odd.

<u>Case 6</u>: The graph spanned by all but vertex $\gamma_1$ of $L_{n,0}$ is the result of removing one vertex from one of the graphs in Figures 12 and 13 (depending on parity of *n*) and attaching a copy of $P_1$. By the previous case and Corollary 4, this is permissible.
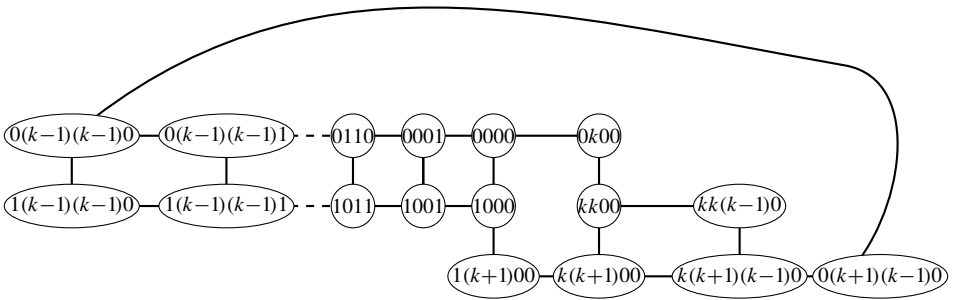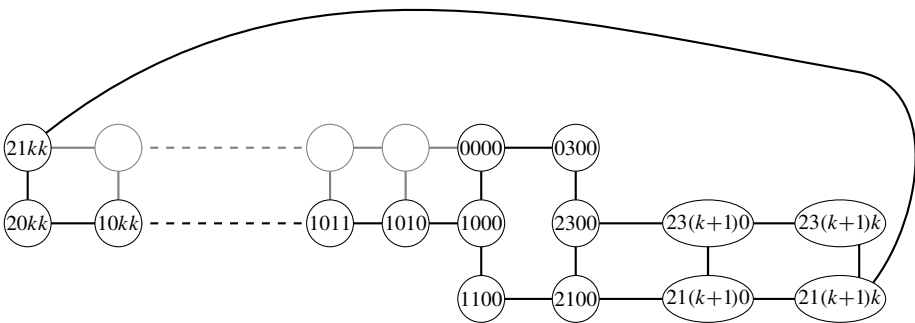
**Figure 13.** Case 5 for $n$ even.



**Figure 14.** Case 7 for $n$ odd.

Case 7: In Figure 14, the colors of the last two vertices are alternately incremented as we move to the left away from the central 6-cycle until the very last step where the color of the first vertex is changed. When a vertex $\alpha_i$ for $1 \le i \le n$ is removed from $L_{n,0}$, there will be one missing vertex among the empty vertices shown in Figure 14. Vertices in positions $\alpha_j$ for $j > i$ should be assigned colorings by changing the second vertex in the coloring below from color 0 to color 1. For the vertices in positions $\alpha_j$ with $j < i$, the colorings should differ from the ones below by changing the first vertex from color from 1 to 0. One can check that the colorings shown in Figure 14, together with the ones just described, span $L_{n,0}$ with vertex $\alpha_i$ removed inside of $\mathcal{C}_{k+2}(I_4)$. The same concept works for $n$ even. The labels are shown in Figure 15. $\qquad \square$

# 4. Future directions

Given time and creativity, it seems certain one could find many other examples of minimal forbidden subgraphs of coloring graphs, but there are several other interesting directions one could explore related to this research. This paper builds on
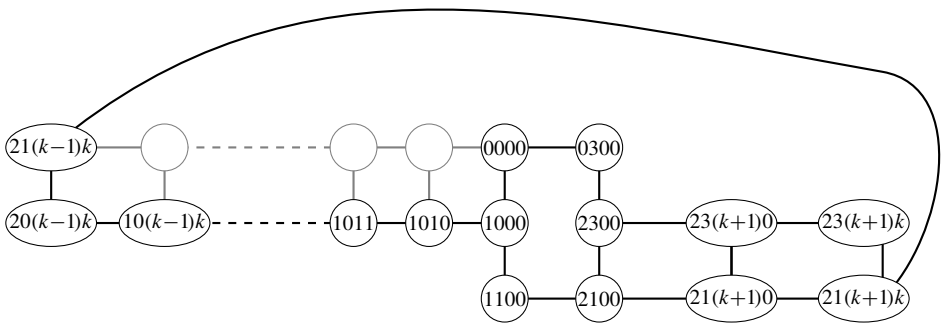
**Figure 15.** Case 7 for *n* even.

[Beier et al. 2016] to provide a template for showing graphs are minimal forbidden, but are there other less brute-force ways to show graphs are minimal forbidden?

Coloring edge labelings are still not completely understood. We provide some necessary conditions for edge labelings of cycles to be proper, but are there others? Are there sufficient conditions for an edge labeling of a cycle to be proper? Closely related to this, can we find a simple method for determining when a partition is an edge labeling partition? Finally, coloring graphs are a particular type of transition graph. To what extent will the methods presented here apply to other types of transition graphs? Other examples of transition graphs can be found in [Cohen and Teicher 2014; Zhang et al. 1988; Haas 2012; Mohar 2007].

## References

[Beier et al. 2016] J. Beier, J. Fierson, R. Haas, H. M. Russell, and K. Shavo, "Classifying coloring graphs", *Discrete Math.* **339**:8 (2016), 2100–2112. MR Zbl

[Cohen and Teicher 2014] M. Cohen and M. Teicher, "Kauffman's clock lattice as a graph of perfect matchings: a formula for its height", *Electron. J. Combin.* **21**:4 (2014), art. id. 4.31. MR Zbl

[Diestel 1997] R. Diestel, *Graph theory*, Graduate Texts in Mathematics **173**, Springer, New York, 1997. MR Zbl

[Dyer et al. 2006] M. Dyer, A. D. Flaxman, A. M. Frieze, and E. Vigoda, "Randomly coloring sparse random graphs with fewer colors than the maximum degree", *Random Structures Algorithms* **29**:4 (2006), 450–465. MR Zbl

[Haas 2012] R. Haas, "The canonical coloring graph of trees and cycles", *Ars Math. Contemp.* **5**:1 (2012), 149–157. MR Zbl

[Jerrum 1995] M. Jerrum, "A very simple algorithm for estimating the number of *k*-colorings of a low-degree graph", *Random Structures Algorithms* **7**:2 (1995), 157–165. MR Zbl

[Mohar 2007] B. Mohar, "Kempe equivalence of colorings", pp. 287–297 in *Graph theory in Paris*, edited by A. Bondy et al., Birkhäuser, Basel, 2007. MR Zbl

[Molloy 2004] M. Molloy, "The Glauber dynamics on colorings of a graph with high girth and maximum degree", *SIAM J. Comput.* **33**:3 (2004), 721–737. MR Zbl

[Vigoda 2000]  E. Vigoda, "Improved bounds for sampling colorings", *J. Math. Phys.* **41**:3 (2000), 1555–1569. MR Zbl

[Zhang et al. 1988]  F. J. Zhang, X. F. Guo, and R. S. Chen, "Z-transformation graphs of perfect matchings of hexagonal systems", *Discrete Math.* **72**:1-3 (1988), 405–415. MR Zbl

falvara2@calstatela.edu            *California State University, Los Angeles, CA, United States*

a_butts1@u.pacific.edu            *University of the Pacific, Stockton, CA, United States*

lafa1550@colorado.edu            *Department of Mathematics, University of Colorado, Boulder, CO, United States*

hrussell@richmond.edu            *Department of Mathematics, University of Richmond, Richmond, VA, United States*

# Computing indicators of Radford algebras

Hao Hu, Xinyi Hu, Linhong Wang and Xingting Wang

(Communicated by Kenneth S. Berenhaut)

We compute higher Frobenius–Schur indicators of Radford algebras in positive characteristic and find minimal polynomials of these linearly recursive sequences. As a result of the work of Kashina, Montgomery and Ng, we obtain gauge invariants for the monoidal categories of representations of Radford algebras.

## 1. Introduction

In group theory, the Frobenius–Schur (FS) indicator provides a criterion, depending on its possible values 1, 0, or −1, for determining whether an irreducible representation of a finite group $G$ is real, complex or quaternionic. This result was generalized to any semisimple Hopf algebra over an algebraically closed field of characteristic zero in [Linchenko and Montgomery 2000]. Kashina, Montgomery and Ng [Kashina et al. 2012] proposed a definition of higher Frobenius–Schur (FS) indicators for an arbitrary finite-dimensional Hopf algebra, which further generalizes the notion given in [Kashina et al. 2006] regarding the regular representation of a semisimple Hopf algebra. Moreover, they proved that these indicators are gauge invariant under gauge equivalence in the sense of [Kassel 1995]. Later, the properties of these indicators were further discussed by Shimizu [2015], who mainly focused on the complex Hopf algebras.

The definition of higher FS indicators of the regular representation of a finite-dimensional Hopf algebra is straightforward by taking the trace of the Sweedler powers followed by the antipode; see [Kashina et al. 2012, Definition 2.1]. But to find their values can be arithmetically challenging over the complex numbers, e.g., in the case of the indicators of Taft algebras; see [Kashina et al. 2012, §3]. Besides Taft algebras, another well-studied Hopf algebra with simple defining relation is the Radford algebra $R(p)$, which was introduced in [Radford 1977, 4.13] and is over a base algebraically closed field of prime characteristic $p$. It was proved in [Wang and Wang 2014] that $R(p)$ is the only noncommutative and noncocommutative pointed Hopf algebra of dimension $p^2$.

In this short note, we find that the higher FS indicators of the Radford algebra $R(p)$ are

$$\{\nu_n(R(p))\}_{n\geq 1} = \{\underbrace{1, \ldots, 1}_{p-1}, 0, \underbrace{1, \ldots, 1}_{p-1}, 0, \ldots\}.$$

Our approach is via concrete computation involving the left integrals of the Radford algebra and those of its dual Hopf algebra. Our result verified, in the case of the Radford algebra, a theorem by Shimizu [2015, Corollary 4.6] on higher FS indicators over positive characteristic, which states that the sequence of indicators always appears periodically in positive characteristic. As a result of the work of Kashina, Montgomery and Ng, we obtain gauge invariants for the monoidal category of the representation of Radford algebras. Moreover, we also find the minimal polynomial of the sequence of indicators of the Radford algebra.

## 2. Preliminaries

Throughout, $\Bbbk$ is an algebraically closed field, $H$ is a finite-dimensional Hopf algebra over $\Bbbk$. We use the standard notation $(H, m, u, \Delta, \varepsilon, S)$, where $m : H \otimes H \to H$ is the multiplication map, $u : \Bbbk \to H$ is the unit map, $\Delta : H \to H \otimes H$ is the comultiplication map, $\varepsilon : H \to \Bbbk$ is the counit map, and $S : H \to H$ is the antipode. The vector space dual of $H$ is also a Hopf algebra and will be denoted by $H^*$. The bialgebra maps and antipode of $H^*$ are given by $(m_{H^*}, u_{H^*}, \Delta_{H^*}, \varepsilon_{H^*}, S_{H^*}) = (\Delta^*, \varepsilon^*, m^*, u^*, S^*)$, where $*$ is the transpose. We use the Sweedler notation $\Delta(h) = \sum h_{(1)} \otimes h_{(2)}$. If $f, g \in H^*$, then $fg(h) = \sum f(h_{(1)})g(h_{(2)})$ for any $h \in H$ and $\varepsilon_{H^*}(f) = f(1)$.

**2.1. Definition** [Montgomery 1993, Definition 2.1.1]. A left integral in $H$ is an element $\Lambda \in H$ such that $h\Lambda = \varepsilon(h)\Lambda$ for all $h \in H$; a right integral in $H$ is an element $\Lambda' \in H$ such that $\Lambda'h = \varepsilon(h)\Lambda'$ for all $h \in H$. The spaces of left and right integrals are denoted by $\int_H^l$ and $\int_H^r$, respectively.

**2.2. Lemma** [Montgomery 1993, Theorem 2.1.3]. *The spaces $\int_H^l$ and $\int_H^r$ are each one-dimensional.*

**2.3. Lemma.** *Suppose $\lambda \in H^*$. Then $\lambda$ is a left integral of $H^*$ if and only if $\sum h_{(1)}\lambda(h_{(2)}) = \lambda(h)$ for any $h \in H$. A similar criterion holds for a right integral of $H^*$, i.e., $\lambda$ is a right integral of $H^*$ if and only if $\sum \lambda(h_{(1)})h_{(2)} = \lambda(h)$ for any $h \in H$.*

*Proof.* By definition, $\lambda$ is a left integral in $H^*$ if and only if $f\lambda = \varepsilon_{H^*}(f)\lambda$ for any linear function $f \in H^*$. That is, $f\lambda(h) = \varepsilon_{H^*}(f)\lambda(h)$ for any $h \in H$. By duality, this is equivalent to $\sum f(h_{(1)})\lambda(h_{(2)}) = f(1)\lambda(h)$ or $f\left(\sum h_{(1)}\lambda(h_{(2)})\right) = f(1\lambda(h))$ since $f$ is linear. Note that $f$ is arbitrary in $H^*$. We have $\lambda$ is a left integral in $H^*$

if and only if $\sum h_{(1)}\lambda(h_{(2)}) = \lambda(h)$ for any $h \in H$. The proof for right integrals is the same. $\qquad\square$

**2.4. Definition** [Kashina et al. 2012, Definition 2.1]. Let $n$ be a positive integer. Suppose $h_1, \ldots, h_n \in H$. Then the $n$-th power of multiplication is defined as

$$m^{(n)}(h_1 \otimes \cdots \otimes h_n) = h_1 \cdots h_n.$$

Let $h \in H$. The $n$-th power of comultiplication is defined to be

$$\Delta^{(n)}(h) = \begin{cases} h, & n = 1, \\ (\Delta^{(n-1)} \otimes \mathrm{id})(\Delta(h)), & n \geq 2. \end{cases}$$

The $n$-th Sweedler power of $h$ is defined to be

$$P_n(h) = h^{[n]} = \begin{cases} \varepsilon(h)1_H, & n = 0, \\ m^{(n)} \circ \Delta^{(n)}(h), & n \geq 1. \end{cases}$$

The $n$-th indicator of $H$ is given by

$$\nu_n(H) = \mathrm{Tr}(S \circ P_{n-1}).$$

In particular, $\nu_1(H) = 1$ and $\nu_2(H) = \mathrm{Tr}(S)$.

Let $H$ and $K$ be two finite-dimensional Hopf algebras over $\Bbbk$ such that the two representation categories $\mathrm{Rep}(H)$ and $\mathrm{Rep}(K)$ are monoidally equivalent. By [Ng and Schauenburg 2008, Theorem 2.2], $H \cong K^F$, where $K^F$ is a Drinfeld twist by a gauge transformation $F$ on $H$ which satisfies some 2-cocycle conditions. Then $H$ and $K$ are said to be *gauge equivalent* Hopf algebras.

**2.5. Theorem** [Kashina et al. 2012, Theorem 2.2, Corollary 2.6]. *The sequence* $\{\nu_n(H)\}$ *is an invariant of the gauge equivalence class of Hopf algebras of $H$; that is, if $H$ and $K$ are gauge equivalent then $\{\nu_n(H)\} = \{\nu_n(K)\}$. Suppose $\lambda \in H^*$ and $\Lambda \in H$ are both left integrals (or both right integrals) such that $\lambda(\Lambda) = 1$. Then*

$$\nu_n(H) = \lambda(\Lambda^{[n]})$$

*for all positive integers $n$.*

**2.6. Proposition** [Shimizu 2015, Corollary 4.6]. *Suppose* $\mathrm{char}\,\Bbbk > 0$. *Then, for any finite-dimensional Hopf algebra $H$ over $\Bbbk$, the sequence $\{\nu_n(H)\}$ is periodic.*

**2.7. Definition.** A sequence $\{a_n\}_{n\geq 1}$ is linearly recursive if there exists a nonzero polynomial $f(x) = f_0 + f_1 x + f_{m-1}x^{m-1} + f_m x^m$ such that

$$f_0 a_n + f_1 a_{n+1} + \cdots + f_m a_{m+n} = 0$$

for any positive integer $n$. In such a case, we say that $\{a_n\}_{n\geq 1}$ satisfies the polynomial $f(x)$. The monic polynomial of the least degree satisfied by a linearly recursive sequence is called the minimal polynomial of the sequence.

**2.8. Proposition** [Kashina et al. 2012, Proposition 2.7]. *The sequence $\{v_n(H)\}$ is linearly recursive and the degree of its minimal polynomial is at most $(\dim H)^2$. The minimal polynomial is also a gauge invariant; that is, if $H$ and $K$ are gauge equivalent, then $\{v_n(H)\}$ and $\{v_n(K)\}$ have the same monic minimal polynomial.*

Next, we consider a free bialgebra $\mathfrak{B}$ and the comultiplication of certain monomials in $\mathfrak{B}$. This information will be used later in our computation of indicators of $R(p)$.

**2.9. Definition.** Let $\mathfrak{B} = \Bbbk\langle g, x\rangle$ be the free $\Bbbk$-algebra on two generators $g$ and $x$. Equipped with the comultiplication and the counit given by

$$\Delta(g) = g \otimes g, \quad \Delta(x) = x \otimes 1 + g \otimes x, \quad \varepsilon(g) = 1 \quad \text{and} \quad \varepsilon(x) = 0,$$

the free algebra becomes the free bialgebra $(\mathfrak{B}, \Delta, \varepsilon)$. Let $C_{k,l}$ denote the sum of all monomials with $k$ $g$'s and $l$ $x$'s, and $C_{0,0} = 1$ and $C_{k,l} = 0$ if $k$ or $l < 0$ by convention.

**2.10. Lemma.** *In the free bialgebra $\mathfrak{B}$, we have*

(a) $C_{k,l} = g\, C_{k-1,l} + x\, C_{k,l-1} = C_{k-1,l}\, g + C_{k,l-1}\, x.$

(b) $\Delta(x^n) = \sum_{k\geq 0} C_{k,n-k} \otimes x^k$ *for $n \geq 0$.*

(c) $\Delta(C_{p,q}) = \sum_{k\geq 0} C_{p+k,q-k} \otimes C_{p,k}.$

*Proof.* Part (a) is clear, since the leftmost (rightmost) factor of any monomial in the sum $C_{k,l}$ is either $g$ or $x$. For (b), we use induction. When $n = 0$,

$$\sum_{k\geq 0} C_{k,n-k} \otimes x^k = C_{0,0} \otimes 1 = 1 \otimes 1 = \Delta(1).$$

When $n = 1$,

$$\sum_{k\geq 0} C_{k,n-k} \otimes x^k = C_{0,1} \otimes 1 + C_{1,0} \otimes x = x \otimes 1 + g \otimes x = \Delta(x).$$

Suppose $\Delta(x^n) = \sum_{k\geq 0} C_{k,n-k} \otimes x^k$. Then

$$\Delta(x^{n+1}) = \Delta(x^n)\Delta(x) = \left(\sum_{k\geq 0} C_{k,n-k} \otimes x^k\right) \cdot (x \otimes 1 + g \otimes x)$$

$$= \sum_{k\geq 0} C_{k,n-k} x \otimes x^k + \sum_{k\geq 1} C_{k-1,n-k+1} g \otimes x^k$$

$$= \sum_{k\geq 0} C_{k,n-k} x \otimes x^k + \sum_{k\geq 1}(C_{k,n-k+1} - C_{k,n-k} x) \otimes x^k$$

$$= \sum_{k\geq 0} C_{k,(n+1)-k} \otimes x^k.$$

To show (c), we use the fact that

$$(\Delta \otimes \mathrm{id})(\Delta(x^n)) = (\mathrm{id} \otimes \Delta)(\Delta(x^n)).$$

By (b), we have

$$(\Delta \otimes \mathrm{id})(\Delta(x^n)) = \sum_{k \geq 0} \Delta(C_{k,n-k}) \otimes x^k = \sum_{p+q=n} \Delta(C_{p,q}) \otimes x^p.$$

On the other hand,

$$(\mathrm{id} \otimes \Delta)(\Delta(x^n)) = \sum_{l \geq 0} C_{l,n-l} \otimes \Delta(x^l) = \sum_{l \geq 0} C_{l,n-l} \otimes \left( \sum_{p \geq 0} C_{p,l-p} \otimes x^p \right)$$

$$= \sum_{p \geq 0} \sum_{l \geq p} C_{l,n-l} \otimes C_{p,l-p} \otimes x^p$$

$$= \sum_{p+q=n} \left( \sum_{l-p=k \geq 0} C_{p+k,q-k} \otimes C_{p,k} \right) \otimes x^p.$$

It then follows that $\Delta(C_{p,q}) = \sum_{k \geq 0} C_{p+k,q-k} \otimes C_{p,k}$. $\qquad \square$

**2.11. Lemma.** *In the free bialgebra* $\mathfrak{B}$, *we have*

$$(g^i x^j)^{[n+1]} = \sum_{0 \leq k_1 + \cdots + k_n \leq j} g^i C_{k_1 + \cdots + k_n, \, j - (k_1 + \cdots + k_n)} g^i C_{k_1 + \cdots + k_{n-1}, k_n} \cdots g^i C_{k_1, k_2} g^i C_{0, k_1}.$$

*Proof.* By induction on $n$, using Lemma 2.10, it is easy to see that

$$\Delta^{(n+1)}(C_{p,q}) = \sum_{0 \leq k_1 + \cdots + k_n \leq q} C_{p+k_1+\cdots+k_n, \, q-(k_1+\cdots+k_n)}$$
$$\otimes C_{p+k_1+\cdots+k_{n-1}, k_n} \otimes \cdots \otimes C_{p+k_1, k_2} \otimes C_{p, k_1}.$$

Therefore, we have

$$(g^i x^j)^{[n+1]} = m^{(n+1)} \big( \Delta^{(n+1)}(g^i) \Delta^{(n+1)}(x^j) \big)$$

$$= m^{(n+1)}(g^i \otimes \cdots \otimes g^i) \left( \sum_{k \geq 0} \Delta^{(n)}(C_{k, j-k}) \otimes x^k \right)$$

$$= \sum_{0 \leq k_1 + \cdots + k_n \leq j} g^i C_{k_1 + \cdots + k_n, \, j-(k_1+\cdots+k_n)} \cdots g^i C_{k_1, k_2} g^i C_{0, k_1}. \qquad \square$$

## 3. Radford algebras

In this section, the base field $\Bbbk$ is algebraically closed of prime characteristic $p$.

**3.1.** The Radford algebra $R(p)$ [1977, 4.13] was first discussed over a base field $\Bbbk$ of prime characteristic $p$, and was proved in [Wang and Wang 2014] to be the only noncommutative and noncocommutative pointed Hopf algebra of dimension $p^2$

over $\Bbbk$. In fact, one can write $R(p)$ as the quotient Hopf algebra $\mathfrak{B}/\mathscr{R}$, where the ideal $\mathscr{R}$ of $\mathfrak{B}$ is generated by

$$g^p - 1, \quad x^p - x, \quad [g, x] - (g^2 - g) \tag{$\mathscr{R}$}$$

if $p > 2$, or

$$g^2 - 1, \quad x^2 - x, \quad [g, x] - (1 - g)$$

if $p = 2$. It is straightforward to check that the Radford algebra $R(p)$ has dimension $p^2$ and the linear basis can be chosen as $\{g^i x^j \mid 0 \le i, j \le p - 1\}$. We denote by $c_{k,l}$ the image of $C_{k,l}$ (the sum of all monomials with $k$ $g$'s and $l$ $x$'s in $\mathfrak{B}$) in $R(p)$ under the projection $\mathfrak{B} \to \mathfrak{B}/\mathscr{R} = R(p)$. It follows from ($\mathscr{R}$) that, for $0 \le k, l \le p - 1$,

$$c_{k,l} = \binom{k+l}{k} g^k x^l + \sum_{\substack{0 \le i \le p-1 \\ 0 \le j \le l-1}} a_{ij} g^i x^j \quad \text{for some } a_{ij} \in \Bbbk. \tag{1}$$

Moreover, the Radford algebra $R(p)$ is self-dual. The dual basis of $(R(p))^*$ to the chosen basis $\{g^i x^j \mid 0 \le i, j \le p - 1\}$ of $R(p)$ is $\{\delta_{g^i x^j} \mid 0 \le i, j \le p - 1\}$, where $\delta_{g^i x^j}$ are characteristic functions, that is,

$$\delta_{g^i x^j}(g^m x^n) = \begin{cases} 1 & \text{if } m = i, \ n = j, \\ 0 & \text{otherwise.} \end{cases}$$

**3.2. Lemma.** *For the Radford algebra $R(p)$, the integral spaces are given by*

$$\int_{R(p)}^l = \Bbbk \left( \sum_{0 \le i \le p-1} g^i \right) \left( \sum_{1 \le i \le p-1} (-1)^i x^i \right),$$

$$\int_{R(p)}^r = \Bbbk \left( \sum_{1 \le i \le p-1} x^i \right) \left( \sum_{0 \le i \le p-1} g^i \right).$$

*For the dual Hopf algebra $(R(p))^*$, the integral spaces are given by*

$$\int_{R(p)^*}^l = \Bbbk \, \delta_{g x^{p-1}} \quad \text{and} \quad \int_{R(p)^*}^r = \Bbbk \, \delta_{x^{p-1}},$$

*Proof.* Note that $\varepsilon(g) = 1$, $\varepsilon(x) = 0$, and $\varepsilon$ is linear. To show that the element $\Lambda = \left( \sum_{0 \le i \le p-1} g^i \right) \left( \sum_{1 \le i \le p-1} (-1)^i x^i \right)$ is a left integral in $R(p)$, it is sufficient to show that $g \Lambda = \Lambda$ and $x \Lambda = 0$. The first equation is obvious. To show the second, one can check that $[x, g^i] = i g^i (1 - g)$. Hence we have

$$\left[ x, \sum_{i=1}^{p-1} g^i \right] = \sum_{i=1}^{p-1} i g^i (1 - g) = \sum_{i=1}^{p-1} i g^i - \sum_{j=2}^{p} (j-1) g^j = g + \sum_{i=2}^{p-1} g^i + g^p = \sum_{i=0}^{p-1} g^i,$$

and so

$$x\Lambda = x\left(\sum_{0\le i\le p-1} g^i\right)\left(\sum_{1\le i\le p-1}(-1)^i x^i\right)$$

$$= \left(\sum_{i=0}^{p-1} g^i\right)(x+1)\left(\sum_{1\le i\le p-1}(-1)^i x^i\right) = \left(\sum_{i=0}^{p-1} g^i\right)(x^p - x) = 0.$$

Therefore, $\Lambda$ is a left integral in $R(p)$.

To show that the characteristic function $\delta_{gx^{p-1}}$ is a left integral in $R(p)^*$, it is sufficient, by Lemma 2.3, to verify that

$$\sum h_{(1)}\delta_{gx^{p-1}}(h_{(2)}) = \delta_{gx^{p-1}}(h) \quad \text{for } h = g^i x^j \in R(p) \text{ with } 0 \le i, j \le p-1.$$

By Lemma 2.10, we have $\Delta(g^i x^j) = (g^i \otimes g^i)\Delta(x^j) = \sum_{k=0}^{j} g^i c_{k,\,j-k} \otimes g^i x^k$. Hence

$$\sum h_{(1)}\delta_{gx^{p-1}}(h_{(2)}) = \sum_{k=0}^{j}\left(g^i c_{k,\,j-k} \cdot \delta_{gx^{p-1}}(g^i x^k)\right)$$

$$= \begin{cases} gc_{p-1,0} = 1 & \text{if } i = 1, \ j = k = p-1, \\ 0 & \text{otherwise.} \end{cases}$$

On the other hand,

$$\delta_{gx^{p-1}}(h) = \delta_{gx^{p-1}}(g^i x^j) = \begin{cases} 1 & \text{if } i = 1, \ j = p-1, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $\delta_{gx^{p-1}}$ is a left integral in $R(p)^*$. The statements on right integrals can be shown similarly. $\qquad\square$

**3.3. Theorem.** *The higher FS indicators of the Radford algebra $R(p)$ are given by*

$$\nu_n(R(p)) = \begin{cases} 1 & \text{if } n \not\equiv 0 \ (\text{mod } p), \\ 0 & \text{if } n \equiv 0 \ (\text{mod } p). \end{cases}$$

*Proof.* By Lemma 3.2, we choose the left integral $\lambda = \delta_{gx^{p-1}}$ of the dual Hopf algebra $(R(p))^*$, and the left integral $\Lambda = \left(\sum_{0\le i\le p-1} g^i\right)\left(\sum_{1\le i\le p-1}(-1)^i x^i\right)$ of $R(p)$. It is clear that $\lambda(\Lambda) = 1$. By Theorem 2.5, we have

$$\nu_{n+1}(R(p)) = \lambda(\Lambda^{[n+1]}) = \delta_{gx^{p-1}}\left(\sum_{0\le i,j\le p-1}(-1)^j (g^i x^j)^{[n+1]}\right).$$

By Lemma 2.11 and (1), one sees that, for any $0 \le i, j \le p-1$,

$$(g^i x^j)^{[n+1]} \in \text{Span}\left(g^k x^l \mid 0 \le k \le p-1, 0 \le l \le j\right).$$

Hence,

$$\nu_{n+1}(R(p)) = \delta_{gx^{p-1}}\left(\sum_{0 \le i \le p-1}(g^i x^{p-1})^{[n+1]}\right).$$

Suppose $k_1, \ldots, k_n$ are nonnegative integers such that $\sum_{i=1}^{n} k_i = m$. Recall that the multinomial coefficients are given by

$$\binom{m}{k_1, \ldots, k_n} := \frac{(m!)}{(k_1!)\cdots(k_n!)}.$$

Assume that $n \ge 1$. Set $k_{n+1} = p - 1 - k_1 - \cdots - k_n$. By Lemma 2.11, we have

$$\sum_{0 \le i \le p-1}(g^i x^{p-1})^{[n+1]}$$

$$= \sum_{\substack{0 \le i \le p-1 \\ 0 \le k_1,\ldots,k_n \le p-1}}\left(g^i c_{k_1+\cdots+k_n,k_{n+1}} g^i c_{k_1+\cdots+k_{n-1},k_n}\cdots g^i c_{k_1,k_2} g^i c_{0,k_1}\right)$$

$$= \sum_{\substack{0 \le i \le p-1 \\ 0 \le k_1,\ldots,k_n \le p-1}}\left(\binom{p-1}{k_1+\cdots+k_n}\binom{k_1+\cdots+k_n}{k_n}\cdots\binom{k_1+k_2}{k_1}\right.$$
$$\left. g^i(g^{k_1+\cdots+k_n}x^{k_{n+1}})g^i(g^{k_1+\cdots+k_{n-1}}x^{k_n})\cdots g^i(g^{k_1}x^{k_2})g^i(x^{k_1})\right)$$

$$= \sum_{\substack{0 \le i \le p-1 \\ 0 \le k_1,\ldots,k_n \le p-1}}\binom{p-1}{k_1,\ldots,k_{n+1}}g^\kappa x^{p-1},$$

where $\kappa = (n+1)i + nk_1 + (n-1)k_2 + \cdots + k_n$. Therefore,

$$\nu_{n+1}(R(p)) = \sum_{0 \le k_1,\ldots,k_{n+1} \le p-1}\binom{p-1}{k_1,\ldots,k_{n+1}}\delta_{gx^{p-1}}\left(\sum_{i=0}^{p-1}g^\kappa x^{p-1}\right).$$

Suppose the indices $k_1, k_2, \ldots, k_n$ are fixed. Then the inner summation of the above equation becomes

$$\sum_{0 \le i \le p-1}g^\kappa x^{p-1} = \begin{cases} p(g^{(nk_1+(n-1)k_2+\cdots+k_n)}x^{p-1}) = 0 & \text{if } p \mid n+1, \\ (1+g+\cdots+g^{p-1})x^{p-1} & \text{if } p \nmid n+1. \end{cases}$$

In a conclusion, by Fermat's little theorem and for $n \ge 1$, we have

$$\nu_{n+1}(R(p)) = \begin{cases} 0 & \text{if } p \mid n+1, \\ \sum_{k_1,\ldots,k_{n+1}}\binom{p-1}{k_1,\ldots,k_{n+1}} = (n+1)^{p-1} = 1 & \text{if } p \nmid n+1. \end{cases}$$

Note that $\nu_1(R(p)) = 1$. Therefore, we showed that

$$\{\nu_n(R(p))\}_{n \ge 1} = \{\underbrace{1,\ldots,1}_{p-1}, 0, \underbrace{1,\ldots,1}_{p-1}, 0, \ldots\}. \qquad \square$$

**3.4. Proposition.** *The minimal polynomial of the sequence $\{\nu_n(R(p))\}$ is*

$$f(x) = x^p - 1.$$

*Proof.* The first $p + 1$ terms of $\{\nu_n(R(p))\}$ are $1, \ldots, 1, 0, 1$. The degree of the minimal polynomial cannot be less than $p$. Otherwise, $\{\nu_n(R(p))\}$ satisfies a polynomial $f(x) = f_0 + f_1 x_1 + \cdots + f_{p-1} x^{p-1}$. Then

$$A[f_0 \, f_1 \, \ldots \, f_{p-1}]^{\mathrm{T}} = 0,$$

where $A$ is the matrix with 0's on the antidiagonal and 1's elsewhere. Note that the determinant of $A$ is $p - 1$ or $-(p - 1)$. This implies that $f_0 = f_1 = \cdots = f_{p-1} = 0$, a contradiction. Hence the degree of the minimal polynomial is at least $p$. One can verify that $\{\nu_n(R(p))\}$ satisfies the polynomial $f(x) = x^p - 1$. □

## Acknowledgements

## References

[Kashina et al. 2006] Y. Kashina, Y. Sommerhäuser, and Y. Zhu, *On higher Frobenius–Schur indicators*, Mem. Amer. Math. Soc. **855**, American Mathematical Society, Providence, RI, 2006. MR Zbl

[Kashina et al. 2012] Y. Kashina, S. Montgomery, and S.-H. Ng, "On the trace of the antipode and higher indicators", *Israel J. Math.* **188**:1 (2012), 57–89. MR Zbl

[Kassel 1995] C. Kassel, *Quantum groups*, Graduate Texts in Mathematics **155**, Springer, New York, 1995. MR Zbl

[Linchenko and Montgomery 2000] V. Linchenko and S. Montgomery, "A Frobenius–Schur theorem for Hopf algebras", *Algebr. Represent. Theory* **3**:4 (2000), 347–355. MR Zbl

[Montgomery 1993] S. Montgomery, *Hopf algebras and their actions on rings*, CBMS Regional Conference Series in Mathematics **82**, American Mathematical Society, Providence, RI, 1993. MR Zbl

[Ng and Schauenburg 2008] S.-H. Ng and P. Schauenburg, "Central invariants and higher indicators for semisimple quasi-Hopf algebras", *Trans. Amer. Math. Soc.* **360**:4 (2008), 1839–1860. MR Zbl

[Radford 1977] D. E. Radford, "Operators on Hopf algebras", *Amer. J. Math.* **99**:1 (1977), 139–158. MR Zbl

[Shimizu 2015] K. Shimizu, "On indicators of Hopf algebras", *Israel J. Math.* **207**:1 (2015), 155–201. MR Zbl

[Wang and Wang 2014] L. Wang and X. Wang, "Classification of pointed Hopf algebras of dimension $p^2$ over any algebraically closed field", *Algebr. Represent. Theory* **17**:4 (2014), 1267–1276. MR Zbl

hah73@pitt.edu                  *Department of Mathematics, University of Pittsburgh,*
                                *Pittsburgh, PA, United States*

xih40@pitt.edu                  *Department of Mathematics, University of Pittsburgh,*
                                *Pittsburgh, PA, United States*

lhwang@pitt.edu                 *Department of Mathematics, University of Pittsburgh,*
                                *Pittsburgh, PA, United States*

xingting@temple.edu             *Department of Mathematics, Temple University,*
                                *Philadelphia, PA, United States*

# Unlinking numbers of links
# with crossing number 10

## Lavinia Bulai

(Communicated by Colin Adams)

We investigate the unlinking numbers of 10-crossing links. We make use of various link invariants and explore their behaviour when crossings are changed. The methods we describe have been used previously to compute unlinking numbers of links with crossing number at most 9. Ultimately, we find the unlinking numbers of all but two of the 287 prime, nonsplit links with crossing number 10.

## 1. Introduction

A *knot* can be thought of as a knotted piece of string with cross-section a single point and ends glued together to form a closed curve. A *link* is a collection of knots, each knot representing a *component* of the link. A *sublink* of a link is the disjoint union of some of its components. Formally, a knot is a smooth isotopy class of embeddings of $S^1$ in $\mathbb{R}^3$ or $S^3$. Similarly, a *link* is a smooth isotopy class of embeddings of a disjoint union of one or more circles in $\mathbb{R}^3$ or $S^3$. A *smooth isotopy* is a smooth map $F : S^1 \sqcup \cdots \sqcup S^1 \times [0, 1] \to \mathbb{R}^3$ together with a family of embeddings $f_t : S^1 \sqcup \cdots \sqcup S^1 \to \mathbb{R}^3$ such that $f_t(x) = F(x, t)$ for all $x \in S^1 \sqcup \cdots \sqcup S^1$ and $t \in [0, 1]$. A link is *trivial* if it is isotopic to the disjoint union of finitely many circles in a plane.

A link is *oriented* if each of its components is assigned an orientation. There are $2^n$ ways to orient a link with $n$ components, by adding an arrow on each knot, pointing in one of two possible directions. A projection of a link onto a plane together with a set of instructions on under-crossings and over-crossings that suffice to reconstruct the original link is referred to as a *link diagram*. We assume the projection is injective, except for some double points. If the crossings are such that one goes under and over alternately when travelling along each component from an arbitrary point back to itself, then the link diagram is said to be *alternating*. This property is illustrated in Figure 1. A link is *alternating* if it admits an alternating
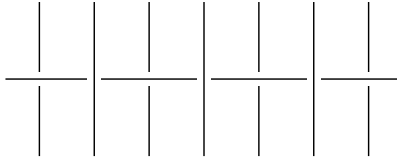
**Figure 1.** The horizontal strand goes alternately over and under the vertical strands.

diagram. A *split link* is a link that has a projection as a disconnected diagram. Otherwise, if every diagram of the link is connected, the link is said to be *nonsplit*. A link $L$ is called *nonprime* if it admits a diagram which is divided into two subsets by a straight line in the plane which intersects the diagram in two points, and such that one does not obtain a diagram of $L$ by replacing either of the subsets by an embedded line segment. A link is *prime* if it is not nonprime.

The *crossing number* of a link is the minimal number of crossings in any of its diagrams. The operation of swapping the two strands that form a crossing such that an under-crossing becomes the over-crossing and vice versa is known as *changing a crossing*. With a sensible choice of crossing changes, one can obtain the trivial link from any given diagram. The *unlinking number* is the minimal number of crossings one has to change in order to obtain the trivial link, where the minimum is taken over all diagrams of the link. In general, unlinking numbers are difficult to determine. In this paper we investigate the unlinking number of each of the 287 prime, nonsplit links with crossing number 10 and at least 2 components by finding constraints on the values it can take. Methods developed by Borodzik, Friedl and Powell [Borodzik et al. 2016], Kauffman and Taylor [1976], Kawauchi [2014], Kohn [1993], Murasugi [1965] and Nagel and Owens [2015] give us lower bounds, whereas upper bounds follow from experiment. Of the links we looked at, the unlinking numbers of two are still unknown and require new techniques to be developed. Good references for basics of knot theory are [Adams 2004; Cromwell 2004; Lickorish 1997; Livingston 1993].

In Section 2 we describe various techniques that can be used to produce lower bounds on unlinking numbers. In Section 3 we give a table of the 10-crossing links and their unlinking numbers, with the exception of two links. For each of these links, we indicate in the table the technique with which the claimed lower bound is produced.

## 2. Lower bounds on unlinking numbers

All the methods we will use throughout this paper to compute unlinking numbers of links with crossing number 10 have previously been used to find unlinking numbers of links with crossing number 9 or less.

We begin with a lemma about real symmetric matrices. The *signature* sign $A$ of a real symmetric matrix $A$ is the number of positive eigenvalues minus the number of negative eigenvalues, counted with multiplicities. The *nullity* of a matrix is the dimension of its kernel.

**Lemma 1.** *Let $A$ be an $n \times n$ real symmetric matrix. Suppose that the matrix $B$ is identical to $A$, apart from one diagonal entry, say $b_{ii} \neq a_{ii}$, where $b_{ii} \in \mathbb{R}$, for some $i \in \{1, \ldots, n\}$. It follows that*:

(i) *The nullity of $B$ differs from the nullity of $A$ by at most $1$.*

(ii) *If $A$ and $B$ have the same nullity and $b_{ii} > a_{ii}$, then the signature of $B$ and the signature of $A$ are related by either* $\text{sign } B = \text{sign } A$ *or* $\text{sign } B = \text{sign } A + 2$.

(iii) *If $A$ and $B$ have different nullities and $b_{ii} > a_{ii}$, then* $\text{sign } B = \text{sign } A + 1$.

*Sketch of proof.* (i) The rank of the matrix $A$ is the dimension of its column space, which in turn is equal to the number of linearly independent columns. By changing the diagonal entry $a_{ii}$ for some $i \in \{1, \ldots, n\}$, the column $i$ will also change; hence the rank of $A$ increases by 1, stays the same, or decreases by 1. However, the change has no effect on the size of $A$. From the rank-nullity theorem it follows that, as the rank changes, the nullity of $A$ will either decrease by 1, stay the same or increase by 1.

(ii)–(iii) The key fact is that by reordering the basis, one can arrange for the leading principal minors of $A$ to form a sequence $d_1, \ldots, d_k, 0, \ldots, 0$, where $d_k \neq 0$ and $k$ is the rank of $A$; furthermore, if $d_i = 0$ then $d_{i-1}d_{i+1} < 0$. This may be done in such a way that the diagonal entry $b_{ii}$ in which $B$ differs from $A$ is either $i = k$ or $i = k+1$. Note that the number of sign changes in this sequence is equal to the number of negative eigenvalues of $A$. See also the proof of Theorem 4 in [Jones 1950]. □

**2.1. *Linking number.*** Let $D$ be a diagram of the oriented link $L$, and $c$ a crossing. There are two possible configurations near $c$, as illustrated in Figure 2. The crossing on the left is said to be *positive*, whereas the crossing on the right is *negative*. Let

$$\epsilon(c) = \begin{cases} 1 & \text{if } c \text{ is a positive crossing,} \\ -1 & \text{if } c \text{ is a negative crossing,} \end{cases}$$
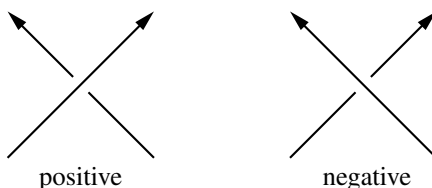


positive        negative

**Figure 2.** Crossing type in an oriented link.

and let $L_1$ and $L_2$ be disjoint sublinks of $L$ such that $L = L_1 \sqcup L_2$. In the diagram of $L$, a crossing may be classified according to the origin of the two strands that form it: $L_1$ with itself, $L_2$ with itself, or $L_1$ with $L_2$. The *linking number* of $L_1$ and $L_2$ is defined as

$$\text{lk}_D(L_1, L_2) = \frac{1}{2} \sum_{c \in \pi(L_1) \cap \pi(L_2)} \epsilon(c),$$

where $\pi(L_i)$ is the projection of $L_i$ to the diagram $D$, and we write $c \in \pi(L_1) \cap \pi(L_2)$ if one of the strands in the crossing belongs to $L_1$ and the other to $L_2$. Once an orientation is fixed, the linking number does not depend on the choice of diagram, so we can refer to it as $\text{lk}(L_1, L_2)$. Thus the linking number is an invariant of the link and the chosen sublinks, and a measure of the number of times one sublink winds around the other.

**Proposition 2** [Kohn 1993, Theorem 1]. *Let $L = L_1 \sqcup L_2$ be an oriented link in $\mathbb{R}^3$, where $L_1$ and $L_2$ are disjoint sublinks of $L$. Then the unlinking number of $L$ satisfies*

$$u(L) \geq u(L_1) + u(L_2) + |\text{lk}(L_1, L_2)|,$$

*where $\text{lk}(L_1, L_2)$ is the linking number of $L_1$ and $L_2$.*

*Proof.* Consider some crossing in a diagram $D$ of the link $L$. If both strands belong to the sublink $L_1$, then changing the crossing will affect neither the sublink $L_2$ nor the linking number of $L_1$ and $L_2$. Similarly, if both strands belong to $L_2$, then changing the crossing will affect neither $L_1$ nor the linking number of the sublinks. However, if one strand belongs to $L_1$ and the other to $L_2$, then changing the crossing will have no effect on the two sublinks, but the linking number will change by 1. Let us now consider an unlinking sequence that realises $u(L)$. The number of crossing changes between $L_1$ and $L_2$ is then bounded below by $|\text{lk}(L_1, L_2)|$, and the number of crossing changes completely in $L_1$ or completely in $L_2$ is bounded below by $u(L_1)$ and $u(L_2)$ respectively, thus proving the inequality. $\square$

To illustrate the application of this method, consider the link $L10n96$, oriented as in Figure 3. Let the sublinks $L_1$ and $L_2$ both be Hopf links — red with blue, and green with purple, respectively. The linking number of $L_1$ and $L_2$ is 3, and it follows from an easy application of Proposition 2 that the unlinking number of a Hopf link is 1, so that $u(L10n96) \geq 5$. Therefore, the link has unlinking number 5, as it can be converted to the trivial link with four components by changing the five crossings indicated in the figure.

**2.2. *Link signature.*** For the next method, let us begin by describing a formula for the signature of a link. Consider a diagram of the link $L$ with chessboard shading, so that no two adjacent regions share the same colour. Assign an *incidence*
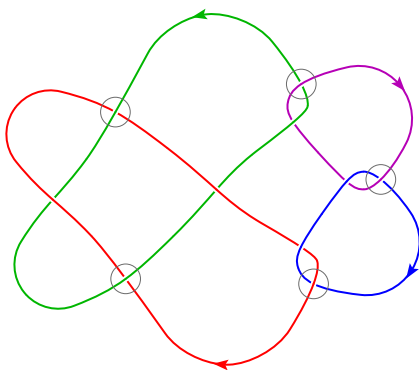
**Figure 3.** One possible way to unlink $L10n96$.

*number* $\iota(c)$ to each crossing in the diagram, by letting

$$\iota(c) = \begin{cases} 1 & \text{if } c \text{ is a right-handed crossing,} \\ -1 & \text{if } c \text{ is a left-handed crossing.} \end{cases}$$

Handedness is illustrated in Figure 4. Note that this is defined using the shading, and is independent of orientation. Let the $n+1$ unshaded regions in the diagram of $L$ be $R_0, R_1, \ldots, R_n$. Construct the square matrix $G' = (g_{ij})$, with entries

$$g_{ij} = \begin{cases} -\sum \iota(c) & \text{if } i \neq j, \text{ summing over crossings } c \text{ incident to both } R_i \text{ and } R_j, \\ -\sum\limits_{k=0, k\neq i}^{k=n} g_{ik} & \text{if } i = j. \end{cases}$$

After deleting the zeroth row and column of $G'$, another matrix is obtained, namely the symmetric square integer *Goeritz matrix* $G$ of the chessboard-shaded link diagram. Let us now orient the link and consider a crossing $c$ in its diagram. If we discard information on under-crossing and over-crossing, then there are two



right-handed          left-handed
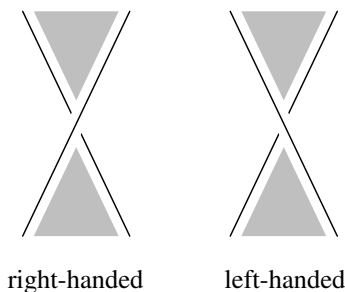
**Figure 4.** Crossings in a chessboard-shaded diagram.
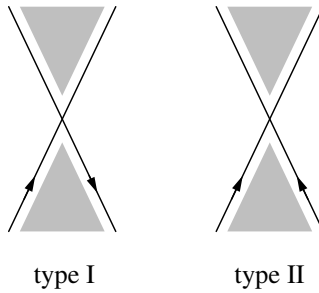
**Figure 5.** Crossings in an oriented chessboard-shaded diagram.

possible configurations near $c$, type I and type II, as illustrated in Figure 5. Define

$$\mu = \sum_{\text{type II}} \iota(c),$$

where the sum is taken over all crossings of type II in the diagram of the link. Then the *signature* of the link is given by

$$\sigma(L) = \operatorname{sign} G - \mu, \qquad\qquad (*)$$

where $\operatorname{sign} G$ is the signature of the Goeritz matrix of the diagram. This definition of signature is due to Gordon and Litherland [1978], who proved it to be equivalent to an older definition using Seifert surfaces. Signature is a link invariant — once an orientation is fixed, the signature remains constant under isotopy. This was proved in [Trotter 1962] for knots and in [Murasugi 1965] for links.

**Proposition 3** [Murasugi 1965, Theorem 10.1; Cochran and Lickorish 1986, Corollary 3.9]. *Let $L$ be an oriented link in $\mathbb{R}^3$. Then the unlinking number of $L$ satisfies*

$$u(L) \geq \tfrac{1}{2}|\sigma(L)|,$$

*where $\sigma(L)$ is the signature of the link.*

*Proof.* Consider the trivial link with $k$ components and the standard diagram consisting of $k$ nonnested circles with no crossings. For one choice of shading, the corresponding Goeritz matrix $G$ of this link is the zero matrix with $k - 1$ rows and columns, which has $\operatorname{sign} G = 0$. Since there are no crossings in this diagram of the link, we have $\mu = 0$. It follows from $(*)$ that the signature of the trivial link is 0, irrespective of the number of components. Now, given an oriented link $L$ with diagram $D$, we aim to obtain the trivial link by changing crossings in $D$. At each step, let $c$ denote the crossing to be changed, and choose the chessboard colouring of the diagram that makes $c$ a double point of type I. Also, relabel the white regions so that $c$ is adjacent to $R_0$ and $R_n$. In the matrix $G'$ of the link, the effect of the crossing change amounts to changing entries $g_{00}$, $g_{0n}$, $g_{n0}$ and $g_{nn}$. Therefore, the
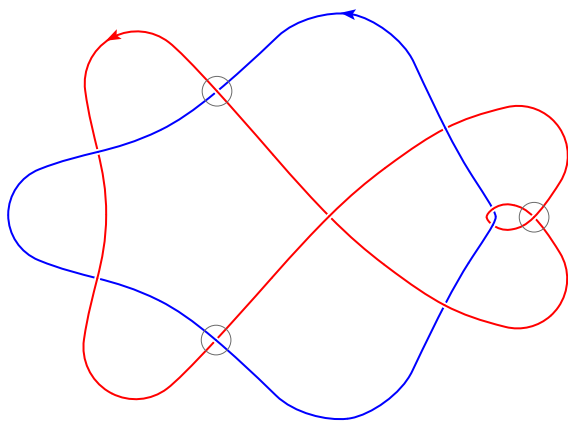
**Figure 6.** One possible way to unlink $L10a99$.

new Goeritz matrix of the link is identical to the original one, except for the diagonal entry $g_{nn}$. By Lemma 1, sign $G$ changes by at most 2. Since $c$ is a double point of type I, changing the crossing will not affect $\mu$. It follows from $(*)$ that $\sigma(L)$, in turn, changes by at most 2. The link is eventually converted to the trivial link, so that its signature changes by at most twice the unlinking number throughout the process, which implies that $|\sigma(L)| \leq 2u(L)$, or equivalently, $u(L) \geq \frac{1}{2}|\sigma(L)|$. $\quad\square$

To illustrate the application of this method, consider the link $L10a99$. Using $(*)$, one may show that the link has signature $-5$ when oriented as in Figure 6, so that $u(L10a99) \geq 3$. Therefore, the link has unlinking number 3, as it can be converted to the trivial link with 2 components by changing the 3 crossings indicated in the figure.

**2.3. *Link determinant and link nullity.*** The *determinant* of a link is defined to be the determinant of its Goeritz matrix. Similarly, the *nullity* of a link is equal to the nullity of its Goeritz matrix, provided that a connected diagram is considered.

**Proposition 4** [Kauffman and Taylor 1976, Corollary 3.21; Kawauchi 2014, Corollary 4.3; Nagel and Owens 2015, Lemma 2.4]. *Let $L$ be a link in $\mathbb{R}^3$, with $k$ components, nullity $\eta(L)$ and determinant $\det L$. Let $u(L)$ be the unlinking number of $L$:*

(a) *Then*
$$u(L) \geq k - 1 - \eta(L).$$

(b) *If $u(L) \leq k - 1$, then $\det L = 2^{k-1}c^2$ for some $c \in \mathbb{Z}$.*

*Proof.* Consider the trivial link with $k$ components and a connected diagram consisting of $k$ circles sitting in a row, with two crossings between each adjacent pair of circles and no other crossings. For either choice of shading, the Goeritz
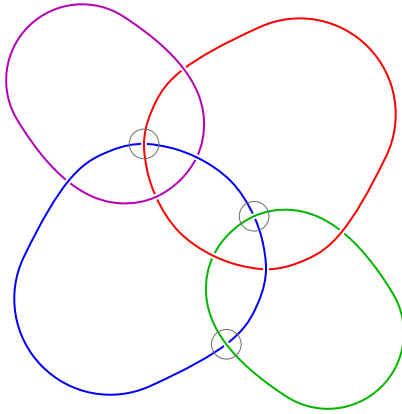
**Figure 7.** One possible way to unlink $L10a169$.

matrix $G$ of this link is the zero matrix with $k - 1$ rows and columns, which has nullity $k - 1$. Now, given a diagram of a link $L$ with $k$ components and nullity $\eta(L)$, construct the matrix $G'$ as in Section 2.2 and change a crossing. As before, we can arrange so that the change affects only one entry in the Goeritz matrix of $L$, namely the bottom right element $g_{nn}$. It follows from Lemma 1 that the nullity of the Goeritz matrix will change by at most 1, and so too will the nullity of the link. Since $L$ is converted to the trivial link with $u(L)$ crossing changes, its nullity cannot change by more than the unlinking number, giving $u(L) \geq |(k-1) - \eta(L)| \geq k - 1 - \eta(L)$. For a proof of part (b) see [Kawauchi 2014], where this statement is shown to follow from a stronger condition involving multivariable Alexander polynomials, or [Nagel and Owens 2015]. $\qquad\square$

To illustrate the application of the method described in Proposition 4(a), consider the link $L10a169$ with four components and nullity 0, so that $u(L10a169) \geq 3$. Therefore, the link has unlinking number 3, as it can be converted to the trivial link with four components by changing the three crossings indicated in Figure 7.

For the method described in part (b), let $L$ be the link $L10n33$, with $k = 2$ components and determinant $\det L = 48$. Suppose that $u(L) \leq 1$. Then by the proposition, $c^2 = 24$ for some $c \in \mathbb{Z}$, a contradiction that gives $u(L) > 1$. Therefore, the link has unlinking number 2, as it can be converted to the trivial link with two components by changing the two crossings indicated in Figure 8.

Every $n \times n$ integer matrix $M$ can be transformed by a finite sequence of row and column operations into a diagonal matrix, whose diagonal entries form a sequence $\{a_1, a_2, \ldots, a_r, 0, \ldots, 0\}$, where $a_i$ is nonnegative and $a_i$ divides $a_{i+1}$. This diagonal matrix is independent of the sequence of row and column operations, and is called the *Smith normal form* of $M$. The matrix $M$ presents the quotient group
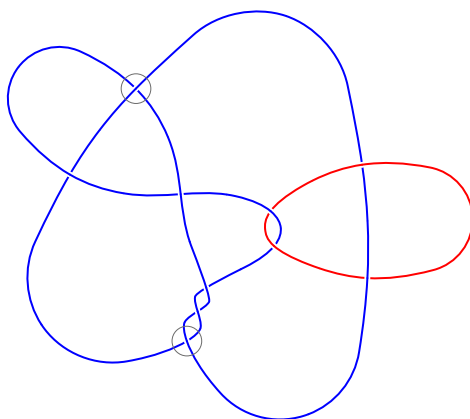
**Figure 8.** One possible way to unlink $L10n33$.

$\mathbb{Z}^n/M\mathbb{Z}^n$, which is cyclic if and only if the Smith normal form $S$ of $M$ satisfies $s_{ii} = 1$ for $i = 1, \ldots, n-1$, and $s_{nn} = \det M$.

**Proposition 5** [Nagel and Owens 2015, Lemma 4.1]. *Let $L$ be a link with two components in $\mathbb{R}^3$ and determinant* $\det L$ *such that its unlinking number satisfies* $u(L) < 3$. *Suppose that the Goeritz matrix of $L$ presents a finite cyclic group. Then at least one of the following statements holds*:

- *$\det L$ is a multiple of 4, and the absolute value of at least one of the signatures of $L$ is 1.*
- *$\det L$ is a multiple of 16.*
- *$\det L = 2t^2$ for some $t \in \mathbb{Z}$.*

The proof of this proposition is based on a 4-dimensional manifold bounded by the double branched cover $Y$ of the link $L$. This gives constraints on the linking form of $Y$, which in turn gives constraints on the determinant and signature of $L$. For details see [Nagel and Owens 2015].

To illustrate the application of this method, let $L$ be the link $L10a54$, with two components and determinant 78. The Smith normal form of the Goeritz matrix $G$ of $L$ is

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 78 \end{bmatrix},$$

so that $G$ presents a finite cyclic group. The determinant of $L$ is neither a multiple of 4, nor a multiple of 16, nor twice the square of some $t \in \mathbb{Z}$, so that $u(L) \geq 3$ by Proposition 5. Therefore, the link has unlinking number 3, as it can be converted to
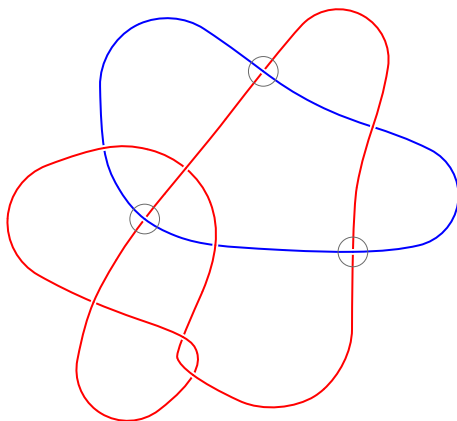
**Figure 9.** One possible way to unlink $L10a54$.

the trivial link with two components by changing the three crossings indicated in
Figure 9.

The following lemma can be viewed as a signed refinement of Proposition 4(a).

**Lemma 6** [Nagel and Owens 2015, Lemma 2.2]. *If an oriented link L with k
components, signature $\sigma(L)$ and nullity $\eta(L)$ is converted to the trivial link by
changing p positive crossings and n negative crossings in some diagram D of the
link, then*

$$p \geq \frac{-\sigma(L) - \eta(L) + k - 1}{2}.$$

*Proof.* Let $c$ be a positive crossing in the diagram of $L$, and choose the chessboard
colouring of $D$ that makes $c$ a double point of type I. In this situation, $c$ has incidence
number $\iota(c) = -1$. Let $G$ be the Goeritz matrix of the diagram and suppose we
change the crossing $c$. As in the proof of Proposition 3, we are free to relabel the
white regions, so that the new Goeritz matrix of the link is identical to the original
one, except for one diagonal entry. After the change, $c$ is still a double point of
type I, but its incidence number becomes $\iota(c) = 1$. Therefore, the diagonal entry
that distinguishes between the two Goeritz matrices increases. By Lemma 1, if
the nullity of $G$ stays the same, then the signature of $G$ either stays the same or
increases by 2, and following $(*)$, so too does $\sigma(L) + \eta(L)$. If the nullity changes,
it can only be by 1, in which case Lemma 1 tells us that the signature of $G$ increases
by 1, and consequently, $\sigma(L) + \eta(L)$ stays the same or increases by 2. By a similar
argument, changing a negative crossing causes $\sigma(L) + \eta(L)$ to either stay constant
or decrease by 2. As we have seen previously, the signature and nullity of the trivial
link with $k$ components add up to $k - 1$. The link $L$ is eventually converted to the
trivial link, so that $\sigma(L) + \eta(L)$ increases by at most twice the number of positive
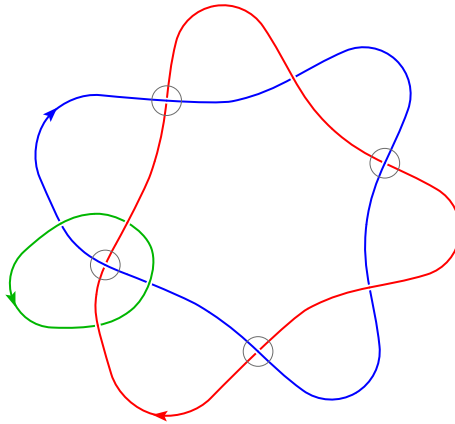
**Figure 10.** One possible way to unlink $L10a138$.

crossings we change, giving $(k-1)-(\sigma(L)+\eta(L)) \le 2p$, or equivalently,

$$p \ge \frac{-\sigma(L)-\eta(L)+k-1}{2}. \qquad \square$$

**2.4. *Lattice embeddings.*** Suppose that the set of vectors $\{a_1, \ldots, a_n\}$ forms a basis for $\mathbb{R}^n$ over $\mathbb{R}$. These vectors span a *lattice* $\Lambda$, which is the set of all linear combinations $\{m_1 a_1 + \cdots + m_n a_n\}$ with $m_i \in \mathbb{Z}$, $i = 1, \ldots, n$. Let $\{b_1, \ldots, b_k\}$ be a set of vectors in $\Lambda$. These vectors span a *sublattice* $\Lambda_b \subset \Lambda$, which is the set of all linear combinations $\{n_1 b_1 + \cdots + n_k b_k\}$ with $n_j \in \mathbb{Z}$, $j = 1, \ldots, k$. The sublattice $\Lambda_b$ of $\Lambda$ is called *primitive* if for all $v \in \Lambda$ and for all $m \in \mathbb{N}$, if $mv \in \Lambda_b$ then $v \in \Lambda_b$. Nagel and Owens gave an obstruction to equality in the lower bound from Lemma 6, which we describe next.

**Proposition 7** [Nagel and Owens 2015, Corollary 3]. *Let $L$ be an oriented non-split alternating link, with $k$ components and signature $\sigma(L)$. Suppose $L$ can be converted to the trivial link by changing $p = \frac{1}{2}(-\sigma(L)+k-1)$ positive crossings and $n$ negative crossings in some diagram of $L$. Let $m$ be the rank of the positive-definite Goeritz matrix $G$ associated to an alternating diagram of $L$, and define $l = m + 2(n+p) - k + 1$. Then $G$ admits a factorisation as $A^T A$, where $A$ is an integer $l \times m$ matrix. Moreover, there exist vectors $v_i$ for $i = 1, \ldots, p+n$ in $(\mathrm{Col}\, A)^\perp \subset \mathbb{Z}^l$ spanning a primitive sublattice of $\mathbb{Z}^l$ such that $v_i \cdot v_j = 2\delta_{ij}$, where $\delta_{ij}$ is the Kronecker delta.*

The proof of Proposition 7 uses results of Gordon and Litherland [1978], as well as the celebrated diagonalisation theorem of Donaldson [1987], and is based on a generalisation of earlier work by Cochran and Lickorish [1986].

To illustrate the application of this method, let $L$ be the link $L10a138$, with three components, determinant 48 and nullity 0. By part (b) of Proposition 4, $u(L) > 2$,

and we aim to obstruct it from being 3. When oriented as in Figure 10, the link has signature $-4$. Suppose $L$ can be converted to the trivial link by changing $p$ positive crossings and $n$ negative crossings in some diagram. By Lemma 6, $p \geq 3$. Thus the only possibility if $u(L) = 3$ is to have $p = 3$ and $n = 0$, which we will show cannot occur. Suppose $p = 3$ and $n = 0$. The positive-definite Goeritz matrix of the chosen alternating diagram is

$$G = \begin{bmatrix} 7 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix},$$

which has rank $m = 3$. Keeping the notation in Proposition 7, we have $l = 7$. For any factorisation of $G$ as $A^T A$, where $A$ is a $7 \times 3$ integer matrix and $A^T$ is its transpose, another may be obtained by interchanging the second and third columns of $A$, permuting the rows of $A$, or multiplying a subset of the rows of $A$ by $-1$. Up to these symmetries, we are left with nine solutions:

$$\begin{bmatrix} -1 & 1 & 1 \\ 2 & 0 & 0 \\ 1 & -1 & 1 \\ -1 & -1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 1 & 1 \\ 2 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ -1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} -2 & 1 & 0 \\ 1 & 0 & 1 \\ -1 & -1 & 1 \\ -1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$\begin{bmatrix} -2 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 2 & 0 & 0 \\ 1 & -1 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} -2 & 1 & 0 \\ -1 & -1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

It is straightforward to check that for any matrix $A$ in this list, there does not exist a set of vectors $\{v_1, v_2, v_3\}$ in the orthogonal complement of the column space of $A$ such that $v_i \cdot v_j = 2\delta_{ij}$, so that $u(L) \geq 4$ by Proposition 7. Therefore, the link has unlinking number 4, as it can be converted to the trivial link with three components by changing the four crossings indicated in Figure 10.

In general, the method based on Proposition 7 gives a somewhat involved algo-rithm to obstruct equality in Lemma 6, leading to improved lower bounds on the unlinking number. All possible factorisations of the Goeritz matrix can be found by hand, but this can also be done using the command *OrthogonalEmbeddings* provided by GAP [2015].
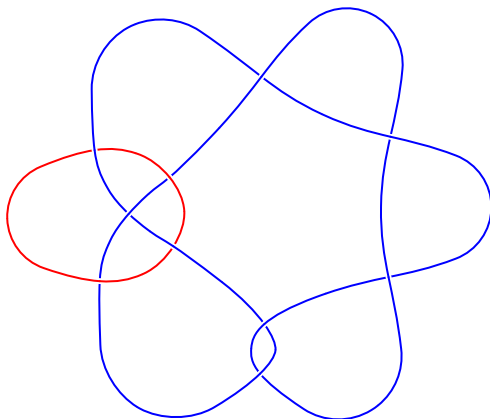
**Figure 11.** Diagram of $L10a7$.

So far, five methods that give lower bounds on the unlinking number of a link — alone or combined — have been described in Propositions 2, 3, 4, 5 and 7. The next method was developed by Kohn [1993].

**2.5. *Covering links.*** Let $p : \mathbb{C} \times \mathbb{R} \to \mathbb{C} \times \mathbb{R}$ be the map taking $(z, t)$ to $(z^2, t)$. Let $L$ be a link with two components, say $L = A \sqcup B$, where $A$ is the trivial knot and $\mathrm{lk}(A, B) = 0$. Assume, after isotopy in $S^3 = \mathbb{R}^3 \cup \{\infty\} = \mathbb{C} \times \mathbb{R} \cup \{\infty\}$, that $A$ is $0 \times \mathbb{R}$, and let $\widetilde{B}$ be the preimage of $B$ under $p$. We refer to $\widetilde{B}$ as the *covering link* of $B$ under $p$.

**Proposition 8** [Kohn 1993, Method 5]. *Let $L$ be a link with two components, say $A$ and $B$, such that $A$ is the trivial knot and $\mathrm{lk}(A, B) = 0$. If $L$ is unlinked by a single crossing change involving $B$ only, then the unlinking number of $\widetilde{B}$ is at most* 2.

*Sketch of proof.* Suppose $L$ can be converted to the trivial link by changing a single crossing $c$, with both strands of $c$ belonging to component $B$. We may isotope $B$ so that it lies near the plane $\mathbb{C} \times \{0\}$ and its projection onto this plane contains the unlinking crossing $c$. The preimage $\widetilde{B}$ of $B$ will then contain two crossings $c_1$ and $c_2$, which are the preimage of $c$ under $p$. Changing $c$ converts $L$ to the unlink; therefore changing $c_1$ and $c_2$ must convert $\widetilde{B}$ to the unlink, since the preimage under $p$ of a circle in $\mathbb{C} \times \{0\}$ not containing the origin is a pair of circles. □

To illustrate the application of this method, let $L$ be the link $L10a7$ shown in Figure 11. The link has two components, namely the red trivial knot $A$ and the blue figure-eight knot $B$, with $\mathrm{lk}(A, B) = 0$. If $L$ can be converted to the unlink by changing a single crossing, then both strands must belong to the knotted component $B$. So suppose that $L$ is converted to the unlink by a single crossing
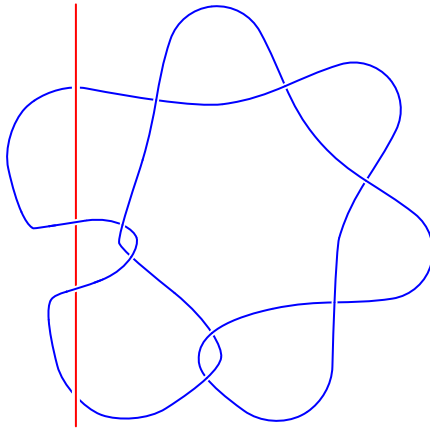
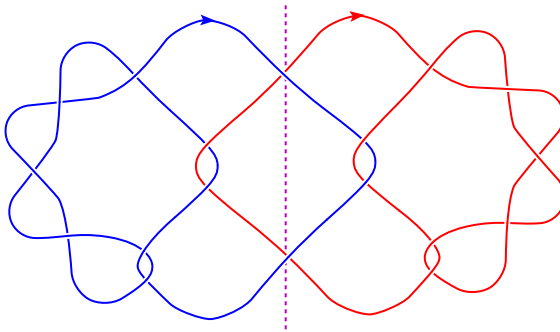**Figure 12.** $L10a7$ when the trivial component is $0 \times \mathbb{R}$.



**Figure 13.** Diagram of $\widetilde{B}$.

change involving $B$ only. After isotopy, assume that $A$ is $0 \times \mathbb{R}$, as depicted in Figure 12.

The preimage of $B$ under the map $p$ is the union of $B$ and its rotated image, glued together to form the covering link $\widetilde{B}$, as in Figure 13. It consists of two stevedore knots — each with unknotting number 1 — with linking number 2 when oriented as shown.

Following Proposition 2, $u(\widetilde{B}) \geq 4$, contradicting Proposition 8. Therefore, $u(L) \geq 2$, and $L$ has unlinking number 2, as it can be converted to the trivial link with two components by changing the two crossings indicated in Figure 14.

## 3. Table of unlinking numbers

Table 1 contains all prime, nonsplit links with crossing number 10 and at least two components, together with the unlinking number $u(L)$ of each link and a proposition
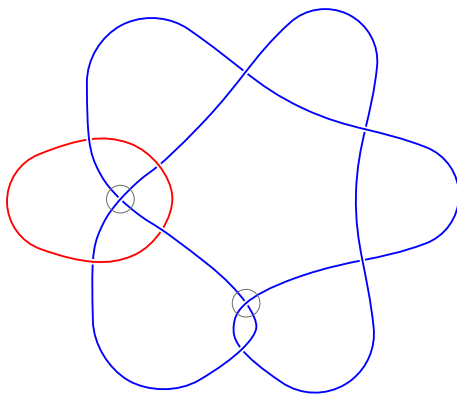
**Figure 14.** One possible way to unlink $L10a7$.

that gives a lower bound that realises $u(L)$. With the exception of $L10n32$ and $L10n34$, the table is complete.

**3.1. _Unknown cases._** Although the methods in this paper were not sufficient to determine the unlinking numbers of two of the links in the table, they still provide partial information. In the following, $p$ is the number of positive crossings and $n$ is the number of negative crossings that we change:

- $L10n32$ has $u(L) \geq 1$ and we conjecture that $u(L) = 2$.
- $L10n34$ has $u(L) \geq 2$ by Proposition 4 and we conjecture that $u(L) = 3$; the cases $p = 0$, $n = 2$ and $p = 1$, $n = 1$ for any choice of orientation are obstructed by Lemma 6 and Proposition 7, respectively.



L10n32                    L10n34
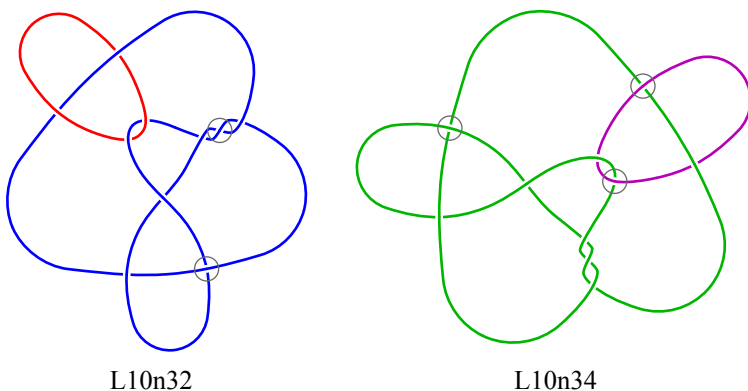
**Figure 15.** Showing a set of crossing changes that unlink the two remaining links.

| link $L$ | $u(L)$ | Prop. | link $L$ | $u(L)$ | Prop. | link $L$ | $u(L)$ | Prop. |
|---|---|---|---|---|---|---|---|---|
| $L10a1$ | 2 | 4b | $L10a48$ | 2 | 2 | $L10a95$ | 1 | 2 |
| $L10a2$ | 2 | 3 | $L10a49$ | 4 | 2 | $L10a96$ | 4 | 2 |
| $L10a3$ | 2 | 8, 2 | $L10a50$ | 3 | 2 | $L10a97$ | 4 | 2 |
| $L10a4$ | 2 | 4b | $L10a51$ | 1 | 2 | $L10a98$ | 4 | 2 |
| $L10a5$ | 2 | 4b | $L10a52$ | 2 | 3 | $L10a99$ | 3 | 3 |
| $L10a6$ | 2 | 4b | $L10a53$ | 1 | 2 | $L10a100$ | 4 | 2 |
| $L10a7$ | 2 | 8, 2 | $L10a54$ | 3 | 5 | $L10a101$ | 4 | 2 |
| $L10a8$ | 3 | 7 | $L10a55$ | 2 | 4b | $L10a102$ | 4 | 2 |
| $L10a9$ | 2 | 3 | $L10a56$ | 2 | 4b | $L10a103$ | 1 | 3 |
| $L10a10$ | 2 | 4b | $L10a57$ | 2 | 2 | $L10a104$ | 2 | 2 |
| $L10a11$ | 3 | 2 | $L10a58$ | 4 | 2 | $L10a105$ | 4 | 2 |
| $L10a12$ | 3 | 2 | $L10a59$ | 2 | 2 | $L10a106$ | 3 | 7 |
| $L10a13$ | 3 | 2 | $L10a60$ | 2 | 2 | $L10a107$ | 4 | 2 |
| $L10a14$ | 2 | 4b | $L10a61$ | 2 | 2 | $L10a108$ | 4 | 2 |
| $L10a15$ | 3 | 2 | $L10a62$ | 3 | 3 | $L10a109$ | 2 | 2 |
| $L10a16$ | 3 | 2 | $L10a63$ | 3 | 5 | $L10a110$ | 4 | 2 |
| $L10a17$ | 3 | 7 | $L10a64$ | 2 | 3 | $L10a111$ | 2 | 4b |
| $L10a18$ | 2 | 2 | $L10a65$ | 2 | 4b | $L10a112$ | 2 | 4b |
| $L10a19$ | 2 | 4b | $L10a66$ | 2 | 2 | $L10a113$ | 3 | 7 |
| $L10a20$ | 2 | 4b | $L10a67$ | 4 | 2 | $L10a114$ | 5 | 2 |
| $L10a21$ | 2 | 4b | $L10a68$ | 2 | 2 | $L10a115$ | 5 | 2 |
| $L10a22$ | 2 | 3 | $L10a69$ | 2 | 2 | $L10a116$ | 5 | 2 |
| $L10a23$ | 3 | 7 | $L10a70$ | 2 | 3 | $L10a117$ | 5 | 2 |
| $L10a24$ | 3 | 7 | $L10a71$ | 2 | 4b | $L10a118$ | 5 | 2 |
| $L10a25$ | 3 | 2 | $L10a72$ | 4 | 2 | $L10a119$ | 5 | 2 |
| $L10a26$ | 3 | 2 | $L10a73$ | 3 | 2 | $L10a120$ | 5 | 2 |
| $L10a27$ | 2 | 3 | $L10a74$ | 4 | 2 | $L10a121$ | 5 | 2 |
| $L10a28$ | 1 | 2 | $L10a75$ | 3 | 2 | $L10a122$ | 4 | 2 |
| $L10a29$ | 1 | 2 | $L10a76$ | 2 | 2 | $L10a123$ | 4 | 2 |
| $L10a30$ | 3 | 2 | $L10a77$ | 4 | 2 | $L10a124$ | 4 | 2 |
| $L10a31$ | 2 | 4b | $L10a78$ | 4 | 2 | $L10a125$ | 4 | 2 |
| $L10a32$ | 2 | 8, 7 | $L10a79$ | 2 | 2 | $L10a126$ | 3 | 2 |
| $L10a33$ | 3 | 2 | $L10a80$ | 2 | 2 | $L10a127$ | 3 | 4b |
| $L10a34$ | 1 | 2 | $L10a81$ | 4 | 2 | $L10a128$ | 3 | 2 |
| $L10a35$ | 2 | 2 | $L10a82$ | 3 | 5 | $L10a129$ | 3 | 2 |
| $L10a36$ | 1 | 2 | $L10a83$ | 3 | 2 | $L10a130$ | 4 | 2 |
| $L10a37$ | 3 | 2 | $L10a84$ | 2 | 2 | $L10a131$ | 4 | 2 |
| $L10a38$ | 4 | 2 | $L10a85$ | 4 | 2 | $L10a132$ | 4 | 2 |
| $L10a39$ | 2 | 2 | $L10a86$ | 2 | 2 | $L10a133$ | 4 | 2 |
| $L10a40$ | 3 | 2 | $L10a87$ | 3 | 2 | $L10a134$ | 4 | 2 |
| $L10a41$ | 2 | 4b | $L10a88$ | 2 | 2 | $L10a135$ | 3 | 2 |
| $L10a42$ | 2 | 2 | $L10a89$ | 2 | 3 | $L10a136$ | 2 | 2 |
| $L10a43$ | 4 | 2 | $L10a90$ | 2 | 4b | $L10a137$ | 4 | 7 |
| $L10a44$ | 4 | 2 | $L10a91$ | 2 | 4b | $L10a138$ | 4 | 7 |
| $L10a45$ | 3 | 2 | $L10a92$ | 2 | 2 | $L10a139$ | 4 | 2 |
| $L10a46$ | 4 | 2 | $L10a93$ | 3 | 7 | $L10a140$ | 2 | 4a |
| $L10a47$ | 3 | 2 | $L10a94$ | 4 | 2 | $L10a141$ | 3 | 7 |

**Table 1.** Unlinking numbers of prime, nonsplit links with crossing number 10.

| link $L$ | $u(L)$ | Prop. | link $L$ | $u(L)$ | Prop. | link $L$ | $u(L)$ | Prop. |
|---|---|---|---|---|---|---|---|---|
| $L10a142$ | 5 | 2 | $L10n17$ | 3 | 2 | $L10n66$ | 4 | 2 |
| $L10a143$ | 5 | 2 | $L10n18$ | 1 | 2 | $L10n67$ | 4 | 2 |
| $L10a144$ | 5 | 2 | $L10n19$ | 3 | 2 | $L10n68$ | 4 | 2 |
| $L10a145$ | 5 | 2 | $L10n20$ | 2 | 4b | $L10n69$ | 4 | 2 |
| $L10a146$ | 5 | 2 | $L10n21$ | 1 | 3 | $L10n70$ | 2 | 2 |
| $L10a147$ | 3 | 2 | $L10n22$ | 1 | 2 | $L10n71$ | 4 | 2 |
| $L10a148$ | 3 | 2 | $L10n23$ | 3 | 3 | $L10n72$ | 4 | 2 |
| $L10a149$ | 3 | 2 | $L10n24$ | 2 | 4b | $L10n73$ | 2 | 2 |
| $L10a150$ | 3 | 2 | $L10n25$ | 3 | 2 | $L10n74$ | 5 | 2 |
| $L10a151$ | 3 | 4b | $L10n26$ | 2 | 2 | $L10n75$ | 5 | 2 |
| $L10a152$ | 5 | 2 | $L10n27$ | 2 | 2 | $L10n76$ | 3 | 2 |
| $L10a153$ | 5 | 2 | $L10n28$ | 3 | 2 | $L10n77$ | 5 | 2 |
| $L10a154$ | 4 | 2 | $L10n29$ | 3 | 2 | $L10n78$ | 5 | 2 |
| $L10a155$ | 4 | 2 | $L10n30$ | 3 | 2 | $L10n79$ | 3 | 2 |
| $L10a156$ | 2 | 2 | $L10n31$ | 3 | 2 | $L10n80$ | 4 | 2 |
| $L10a157$ | 4 | 7 | $L10n32$ | [1, 2] | – | $L10n81$ | 5 | 2 |
| $L10a158$ | 4 | 7 | $L10n33$ | 2 | 4b | $L10n82$ | 5 | 2 |
| $L10a159$ | 5 | 2 | $L10n34$ | [2, 3] | 4b | $L10n83$ | 3 | 2 |
| $L10a160$ | 5 | 2 | $L10n35$ | 2 | 2 | $L10n84$ | 5 | 2 |
| $L10a161$ | 5 | 2 | $L10n36$ | 2 | 2 | $L10n85$ | 3 | 2 |
| $L10a162$ | 3 | 2 | $L10n37$ | 4 | 2 | $L10n86$ | 3 | 2 |
| $L10a163$ | 3 | 4b | $L10n38$ | 4 | 2 | $L10n87$ | 5 | 2 |
| $L10a164$ | 5 | 2 | $L10n39$ | 3 | 3 | $L10n88$ | 4 | 2 |
| $L10a165$ | 4 | 2 | $L10n40$ | 2 | 2 | $L10n89$ | 4 | 2 |
| $L10a166$ | 5 | 2 | $L10n41$ | 2 | 4b | $L10n90$ | 3 | 2 |
| $L10a167$ | 5 | 2 | $L10n42$ | 3 | 3 | $L10n91$ | 4 | 2 |
| $L10a168$ | 5 | 2 | $L10n43$ | 2 | 2 | $L10n92$ | 5 | 2 |
| $L10a169$ | 3 | 3 | $L10n44$ | 1 | 2 | $L10n93$ | 5 | 2 |
| $L10a170$ | 4 | 2 | $L10n45$ | 2 | 2 | $L10n94$ | 5 | 2 |
| $L10a171$ | 5 | 2 | $L10n46$ | 4 | 2 | $L10n95$ | 5 | 2 |
| $L10a172$ | 5 | 2 | $L10n47$ | 4 | 2 | $L10n96$ | 5 | 2 |
| $L10a173$ | 5 | 2 | $L10n48$ | 2 | 2 | $L10n97$ | 5 | 2 |
| $L10a174$ | 5 | 2 | $L10n49$ | 4 | 2 | $L10n98$ | 5 | 2 |
| $L10n1$ | 3 | 2 | $L10n50$ | 3 | 5 | $L10n99$ | 5 | 2 |
| $L10n2$ | 1 | 2 | $L10n51$ | 4 | 2 | $L10n100$ | 3 | 2 |
| $L10n3$ | 2 | 4b | $L10n52$ | 2 | 2 | $L10n101$ | 5 | 2 |
| $L10n4$ | 3 | 2 | $L10n53$ | 3 | 2 | $L10n102$ | 5 | 2 |
| $L10n5$ | 2 | 3 | $L10n54$ | 3 | 3 | $L10n103$ | 3 | 2 |
| $L10n6$ | 2 | 4b | $L10n55$ | 4 | 2 | $L10n104$ | 5 | 2 |
| $L10n7$ | 3 | 2 | $L10n56$ | 1 | 3 | $L10n105$ | 5 | 2 |
| $L10n8$ | 2 | 4b | $L10n57$ | 1 | 4a | $L10n106$ | 4 | 2 |
| $L10n9$ | 1 | 2 | $L10n58$ | 2 | 2 | $L10n107$ | 2 | 2 |
| $L10n10$ | 3 | 2 | $L10n59$ | 2 | 2 | $L10n108$ | 5 | 2 |
| $L10n11$ | 1 | 2 | $L10n60$ | 4 | 2 | $L10n109$ | 5 | 2 |
| $L10n12$ | 2 | 4b | $L10n61$ | 4 | 2 | $L10n110$ | 5 | 2 |
| $L10n13$ | 3 | 2 | $L10n62$ | 3 | 3 | $L10n111$ | 5 | 2 |
| $L10n14$ | 1 | 2 | $L10n63$ | 3 | 7 | $L10n112$ | 5 | 2 |
| $L10n15$ | 3 | 5 | $L10n64$ | 2 | 3 | $L10n113$ | 5 | 2 |
| $L10n16$ | 2 | 2 | $L10n65$ | 4 | 2 | | | |

**Table 1 cont.**

## Acknowledgements

## References

[Adams 2004]  C. C. Adams, *The knot book*, American Mathematical Society, Providence, RI, 2004. MR Zbl

[Bar-Natan and Morrison 2015]  D. Bar-Natan and S. Morrison, "The knot atlas", website, 2015, http://katlas.org.

[Borodzik et al. 2016]  M. Borodzik, S. Friedl, and M. Powell, "Blanchfield forms and Gordian distance", *J. Math. Soc. Japan* **68**:3 (2016), 1047–1080. MR Zbl

[Cha and Livingston 2017a]  J. C. Cha and C. Livingston, "KnotInfo: table of knot invariants", website, 2017, http://www.indiana.edu/~knotinfo.

[Cha and Livingston 2017b]  J. C. Cha and C. Livingston, "LinkInfo: table of link invariants", website, 2017, http://www.indiana.edu/~linkinfo.

[Cochran and Lickorish 1986]  T. D. Cochran and W. B. R. Lickorish, "Unknotting information from 4-manifolds", *Trans. Amer. Math. Soc.* **297**:1 (1986), 125–142. MR Zbl

[Cromwell 2004]  P. R. Cromwell, *Knots and links*, Cambridge University Press, 2004. MR Zbl

[Donaldson 1987]  S. K. Donaldson, "The orientation of Yang–Mills moduli spaces and 4-manifold topology", *J. Differential Geom.* **26**:3 (1987), 397–428. MR Zbl

[GAP 2015]  The GAP Group, "GAP: groups, algorithms, and programming", software, 2015, http://www.gap-system.org. Version 4.7.8.

[Gordon and Litherland 1978]  C. M. Gordon and R. A. Litherland, "On the signature of a link", *Invent. Math.* **47**:1 (1978), 53–69. MR Zbl

[Jones 1950]  B. W. Jones, *The arithmetic theory of quadratic forms*, Carcus Monograph Series **10**, The Mathematical Association of America, Buffalo, NY, 1950. MR Zbl

[Kauffman and Taylor 1976]  L. H. Kauffman and L. R. Taylor, "Signature of links", *Trans. Amer. Math. Soc.* **216** (1976), 351–365. MR Zbl

[Kawauchi 2014]  A. Kawauchi, "The Alexander polynomials of immersed concordant links", *Bol. Soc. Mat. Mex.* (3) **20**:2 (2014), 559–578. MR Zbl

[Kohn 1993]  P. Kohn, "Unlinking two component links", *Osaka J. Math.* **30**:4 (1993), 741–752. MR Zbl

[Lickorish 1997]  W. B. R. Lickorish, *An introduction to knot theory*, Graduate Texts in Mathematics **175**, Springer, 1997. MR Zbl

[Livingston 1993]  C. Livingston, *Knot theory*, Carus Mathematical Monographs **24**, Mathematical Association of America, Washington, DC, 1993. MR Zbl

[Murasugi 1965]  K. Murasugi, "On a certain numerical invariant of link types", *Trans. Amer. Math. Soc.* **117** (1965), 387–422. MR Zbl

[Nagel and Owens 2015] M. Nagel and B. Owens, "Unlinking information from 4-manifolds", *Bull. Lond. Math. Soc.* **47**:6 (2015), 964–979. MR Zbl

[Swenton 2011] F. Swenton, "Kirby calculator", software, 2011, http://kirbycalculator.net.

[Trotter 1962] H. F. Trotter, "Homology of group systems with applications to knot theory", *Ann. of Math.* (2) **76** (1962), 464–498. MR Zbl

lavinia_bulai@yahoo.com          *School of Mathematics and Statistics, University of Glasgow, Glasgow, United Kingdom*

# On a connection between local rings and their associated graded algebras

Justin Hoffmeier and Jiyoon Lee

(Communicated by Vadim Ponomarenko)

We study a class of local rings and a local adaptation of the homogeneous property for graded rings. While the rings of interest satisfy the property in the local case, we show that their associated graded $k$-algebras do not satisfy the property in the graded case.

## 1. Introduction and preliminaries

Let $Q = k[\![X_1, X_2, \ldots, X_n]\!]$ denote the power series ring in $n$ variables over the field $k$. Let $J$ be an ideal in $Q$. For an element $b \in J$, the initial form of $b$ is the homogeneous finite sum of lowest-degree terms of $b$, denoted by $b^*$. Let $Q^g = k[X_1, X_2, \ldots, X_n]$ denote the polynomial ring in $n$ variables over the field $k$. The initial ideal of $J$ is the ideal in $Q^g$ generated by all of the initial forms of $J$ and is denoted by $\mathrm{In}(J)$. That is,

$$\mathrm{In}(J) = \left\{ \sum_{i=1}^m a_i b_i^* \mid a_i \in Q^g, \ b_i \in J, \ 1 \le i \le m \right\}.$$

Computations in $\mathrm{In}(J)$ are not always straightforward. The following example is intended to help illustrate some of the nuances of $\mathrm{In}(J)$.

**Example 1.1.** Let $Q = k[\![X, Y]\!]$ and $J = (x^2 + y^3, xy)$. Since

$$(x^2 + y^3)(-x^2 y + x^4 y^5 + x^{13} + \cdots) \quad \text{and} \quad xy(x^3 + xy^3 - x^5 y^4 + \cdots)$$

are in $J$, we have that the initial form of their sum

$$(-x^4 y + x^4 y + x^2 y^4 - x^2 y^4 + x^6 y^5 - x^6 y^5 + x^4 y^8 + x^{15} + x^{13} y^3 + \cdots)^* = x^4 y^8$$

is in $\mathrm{In}(J)$.

Describing $\mathrm{In}(J)$ is not as simple as finding the initial forms of the generators of $J$. The next example is adapted from [Eisenbud 1995], although similar examples can be found in several other texts.

**Example 1.2.** Let $Q = k[\![X, Y]\!]$ and $J = (x^2 + y^3, xy)$. Then $(x^2 + y^3)^* = x^2$ and $(xy)^* = xy$, but $\text{In}(J) = (x^2, y^4, xy)$. In Lemma 2.5, we provide a method to prove this fact for a more general class of rings.

Let $R$ be a commutative local ring with maximal ideal $\mathfrak{m}$ and residue field $k$. By the Cohen structure theorem, the completion of any local ring can be written as a quotient of a regular local ring by an ideal. Hence, if $R$ is a complete local ring then $R = Q/J$, where $J \subseteq (X_1, X_2, \ldots, X_n)^2$.

**Definition 1.3.** Let $R$ be a complete local ring with a minimal Cohen presentation $R = Q/J$, where $J = (f_1, f_2, \ldots, f_l)$ with $f_i \in Q$ for $1 \le i \le l$. If $f_i^*$ has degree $t$ for each $i$ then $R$ is $t$-homogeneous.

In [Hoffmeier and Şega 2017] the authors give a more general version of the above definition. They go on to show that knowing a ring is $t$-homogeneous is helpful for identifying various homological properties. Indeed, Theorem 2.5 of that paper establishes that the $t$-homogeneous property plays an important role connecting these homological traits of local rings.

Let $J = (f_1, f_2, \ldots, f_l) \subseteq Q^{\mathrm{g}}$ be the ideal generated by polynomials $f_i$ in $Q^{\mathrm{g}}$ for $1 \le i \le l$. If each of the $f_i$ is homogeneous of degree $t$ then the quotient $R = Q^{\mathrm{g}}/J$ is a $t$-homogeneous graded $k$-algebra.

The associated graded ring of $R$ with respect to the maximal ideal is the direct sum

$$R^{\mathrm{g}} = \bigoplus_{i \ge 0} \mathfrak{m}^i / \mathfrak{m}^{i+1}.$$

This notation is consistent with $Q^{\mathrm{g}}$. That is, for the local ring $Q = k[\![X_1, X_2, \ldots, X_n]\!]$, we have $Q^{\mathrm{g}} = k[X_1, X_2, \ldots, X_n]$. Furthermore, if $R = Q/J$ then $R^{\mathrm{g}} = Q^{\mathrm{g}}/\text{In}(J)$.

We now state [Hoffmeier and Şega 2017, Lemma 1.3], which also provides further motivation for the terminology given in Definition 1.3.

**Lemma 1.4.** *Let $R$ be a complete local ring. If $R^{\mathrm{g}}$ is a $t$-homogeneous $k$-algebra, then $R$ is a $t$-homogeneous local ring.*

Hoffmeier and Şega [2017, Remark 1.4] also provide a counterexample to show that the converse of the lemma does not hold. We now reproduce this example.

**Example 1.5.** Let $Q = k[\![X, Y]\!]$, $J = (x^2 + y^3, xy)$, and $R = Q/J$. Then $R$ is 2-homogeneous. However, $R^{\mathrm{g}} = Q^{\mathrm{g}}/\text{In}(J) = k[X, Y]/(x^2, y^4, xy)$, which is not 2-homogeneous.

It is significant that the converse of Lemma 1.4 does not hold. Otherwise, the $t$-homogeneous property of a local ring $R$ would depend only on its associated graded $k$-algebra $R^{\mathrm{g}}$, making the connections between the homological properties of $R$ alluded to above (stated in [Hoffmeier and Şega 2017, Theorem 2.5]) also related to $R^{\mathrm{g}}$. The main goal of this note is to identify a larger class of rings for

which the converse of the lemma fails, which consequently further distinguishes the homological nature of local rings from properties of their associated graded $k$-algebras. We achieve this in the next section by generalizing Example 1.5.

Further motivation for our result is the fact that Example 1.5 is stated without proof in [Hoffmeier and Şega 2017] and is therefore further explained by the proof of our more general result.

**Remark 1.6.** Connections between a local ring and its associated graded algebra have been well documented throughout the literature of commutative algebra. For example, if $R^g$ is Cohen–Macaulay then $R$ is Cohen–Macaulay and if $R^g$ is Gorenstein then $R$ is Gorenstein; see, e.g., [Achilles and Avramov 1982]. The text [Bruns and Herzog 1993] also states several of these results and is a good reference for other topics that appear in this note. In his survey on the subject, Fröberg [1987] states that "A local ring is at least as nice as its associated graded ring." Our results provide another example that makes the inequality Fröberg alludes to strict.

## 2. Unassociated $t$-homogeneous local rings

In this section we prove our main result. We begin with a definition.

**Definition 2.1.** Let $R$ be a $t$-homogeneous local ring. If $R^g$ is not a $t$-homogeneous graded $k$-algebra then we say that $R$ is unassociated $t$-homogeneous.

**Theorem 2.2.** *Let* $J = (x^2 + y^t, xy) \subseteq Q = k[\![X, Y]\!]$ *with* $t \geq 3$ *and set* $R = Q/J$. *Then* $R$ *is unassociated* 2*-homogeneous.*

**Remark 2.3.** Note that by setting $t = 3$ in Theorem 2.2, we recover the result in Example 1.5.

We now provide two lemmas which will be used in the proof of the theorem.

**Lemma 2.4.** *Let* $J = (x^2 + y^t, xy) \subseteq Q = k[\![X, Y]\!]$ *with* $t \geq 3$. *Then* $y^t$ *is not in* $\mathrm{In}(J)$.

*Proof.* Suppose $y^t \in \mathrm{In}(J)$. Then

$$y^t = \sum_{i=1}^m a_i b_i^*,$$

where $a_i \in Q^g$, $b_i \in J$, and $1 \leq i \leq m$. For each $i$, let $b_i = c_i(x^2 + y^t) + d_i(xy)$ with $c_i, d_i \in Q$. Hence,

$$y^t = \sum_{i=1}^m a_i (c_i(x^2 + y^t) + d_i(xy))^*.$$

Since the sum equals $y^t$, the terms of the sum that are factors of $xy$ either cancel or are dropped by taking the lowest-degree terms. Therefore,

$$y^t = \sum_{i=1}^m a_i (c_i(x^2 + y^t))^*.$$

Since $t \geq 3$, we have $(c_i(x^2 + y^t))^* = c_i^* x^2$ for each $i$, where $c_i^*$ is the finite sum of lowest-degree terms of $c_i$. Hence

$$y^t = \sum_{i=1}^m a_i c_i^* x^2,$$

which is a contradiction.                                                                 □

**Lemma 2.5.** *Let* $J = (x^2 + y^t, xy) \subseteq Q = k[\![X, Y]\!]$ *as in Lemma 2.4. Then* $\mathrm{In}(J) = (x^2, y^{t+1}, xy)$.

*Proof.* First, we show that $(x^2, y^{t+1}, xy) \subseteq \mathrm{In}(J)$. It is sufficient to show that $x^2, y^{t+1}, xy \in \mathrm{In}(J)$, which is clear since

$$x^2 = (x^2)^*, \quad xy = (xy)^*, \quad y^{t+1} = (y(x^2 + y^t) - x(xy))^*.$$

Next, we show that $\mathrm{In}(J) \subseteq (x^2, y^{t+1}, xy)$. Let $g \in \mathrm{In}(J)$. Then

$$g = a_1 F_1^* + a_2 F_2^* + \cdots + a_n F_n^*,$$

where $a_i \in k[X, Y]$ and $F_i \in J$ for $1 \leq i \leq n$. Therefore, it suffices to show if $F \in J$ then $F^* \in (x^2, y^{t+1}, xy)$. Let $\alpha, \beta \in k[\![X, Y]\!]$ such that $F = \alpha(x^2 + y^t) + \beta xy$. Then

$$F^* = (\alpha x^2 + \alpha y^t + \beta xy)^* = ax^2 + by^t + cxy$$

for some $a, b, c \in k[X, Y]$. If $b = 0$ then $F^* = ax^2 + cxy \in (x^2, y^{t+1}, xy)$.

Assume $b \neq 0$. Since $F^*$ is homogeneous, $b$ is homogeneous and may be written as $b = px^n + qy$, where $p \in k$, $q \in k[X, Y]$, and $n$ is a nonnegative integer. Therefore,

$$F^* = ax^2 + cxy + px^n y^t + qy^{t+1}.$$

If $p = 0$ then we again have the needed form.

Assume $p \neq 0$ and consider two cases for $n$.

<u>Case (i)</u>: Assume $n \geq 1$. Then $px^n y^t = px^{n-1} y^{t-1}(xy)$. Hence,

$$F^* = ax^2 + qy^{t+1} + (c + px^{n-1} y^{t-1})xy$$

has the needed form.

<u>Case (ii)</u>: Assume $n = 0$. Since we have already shown $(x^2, y^{t+1}, xy) \subseteq \mathrm{In}(J)$, we have

$$F^* - ax^2 - qy^{t+1} - cxy = px^n y^t \in \mathrm{In}(J).$$

Since $n = 0$ and $p^{-1} \in k$ we have $y^t \in \mathrm{In}(J)$, which contradicts Lemma 2.4. Therefore, case (ii) does not occur.                                                 □

**Remark 2.6.** A common approach to working with $\mathrm{In}(J)$ is to invoke the use of Gröbner bases. However, we opt for the more elementary method presented above.

We are now ready to prove the theorem.

*Proof of Theorem 2.2.* Since $R$ is artinian, it is complete. Since $(x^2 + y^t)^* = x^2$ and $(xy)^* = xy$, we know $R$ is 2-homogeneous. As noted above,

$$R^{\mathrm{g}} = Q^{\mathrm{g}} / \operatorname{In}(J).$$

By Lemma 2.5, $\operatorname{In}(J) = (x^2, y^{t+1}, xy)$. Hence, $R^{\mathrm{g}}$ is not a graded $k$-algebra and the theorem follows. $\qquad\square$

## References

[Achilles and Avramov 1982] R. Achilles and L. L. Avramov, "Relations between properties of a ring and of its associated graded ring", pp. 5–29 in *Seminar D. Eisenbud/B. Singh/W. Vogel, II*, Teubner-Texte Math. **48**, Teubner, Leipzig, 1982. MR Zbl

[Bruns and Herzog 1993] W. Bruns and J. Herzog, *Cohen–Macaulay rings*, Cambridge Studies in Advanced Mathematics **39**, Cambridge Univ. Press, 1993. MR Zbl

[Eisenbud 1995] D. Eisenbud, *Commutative algebra with a view toward algebraic geometry*, Graduate Texts in Mathematics **150**, Springer, 1995. MR Zbl

[Fröberg 1987] R. Fröberg, "Connections between a local ring and its associated graded ring", *J. Algebra* **111**:2 (1987), 300–305. MR Zbl

[Hoffmeier and Şega 2017] J. Hoffmeier and L. M. Şega, "Conditions for the Yoneda algebra of a local ring to be generated in low degrees", *J. Pure Appl. Algebra* **221**:2 (2017), 304–315. MR Zbl

jhoff@nwmissouri.edu          *Department of Mathematics and Statistics, Northwest Missouri State University, Maryville, MO, United States*

ji.lee@berkeley.edu           *Missouri Academy of Science, Mathematics and Computing, Maryville, MO, 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.