A Regular Expression Language for the Collatz Graph

Jon Seymour
jon@wildducktheories.com
Sydney, Australia

July 2026

Abstract

The Syracuse map $S(n) = (3n+1)/2^{v_2(3n+1)}$ partitions every orbit into *Steiner circuits*: maximal runs of the form $(\mathrm{OE})^*\,\mathrm{OEE}^+$, where each step is classified by the 2-adic valuation $v = v_2(3n+1)$ and equivalently by the residue of the input $n \bmod 8$. The residue alphabet $\{1, 3, 5, 7\}$ encodes the four step types: OE steps ($v = 1$, $n \equiv 3$ or $7 \bmod 8$), OEE steps ($v = 2$, $n \equiv 1 \bmod 8$), and OEEE^+ steps ($v \geq 3$, $n \equiv 5 \bmod 8$). Within this alphabet, every Steiner circuit is conjectured to match the regular expression $(7^*3)?(1|5)$ and every complete Collatz orbit to match $((7^*3)?(1|5))^*$. This paper develops the combinatorial infrastructure supporting these conjectures: the mod-8 step taxonomy, the Steiner circuit decomposition (proved), the 7-run length theorem (proved), and the *mod-24 universe* — the smallest modular arena in which the parity prefix of every odd residue class is fully determined (established by explicit enumeration). The regular-expression characterisation itself is stated as a conjecture; its proof is deferred to the companion paper [2]. Further analytic tools and cycle-theoretic consequences are developed in the companion papers [2, 3].

Status. This is a working paper. All theorems carry complete proofs. Results labelled *Conjecture* are stated precisely but not yet proved. Deferred material is clearly flagged.

Contents

Note on paper numbering

This paper is Paper 1 in a four-paper series. Papers are numbered by their mod-8 significance: Paper 1 (architecture), Paper 5 (tools), Paper 8 (results), Paper 0 (capstone). On the filesystem the series lives under `papers/{32+p}-{slug}/`.

---

1 Introduction

1.1 The Collatz Maps

Let $\mathbb{Z}^+ = \{1, 2, 3, \ldots\}$ be the positive integers, $\mathbb{O} = \{n \in \mathbb{Z}^+ : n \text{ odd}\}$ the positive odd integers, and $v_2(m)$ the 2-adic valuation of m (the largest power of 2 dividing m).

Definition 1.1 (Natural, Syracuse, and Terras maps)**.** *The* natural Collatz map $C : \mathbb{Z}^+ \to \mathbb{Z}^+$ *is defined by*

$$C(n) = \begin{cases} n/2 & \text{if } n \text{ is even,} \\ 3n+1 & \text{if } n \text{ is odd.} \end{cases}$$

The Syracuse map $S : \mathbb{O} \to \mathbb{O}$ *is obtained by collapsing every run of even steps into a single transition:*

$$S(n) = \frac{3n+1}{2^{v_2(3n+1)}}, \qquad n \in \mathbb{O}.$$

Every orbit of C visits $\mathbb{O}$ after at most finitely many even steps, so S records exactly the odd nodes visited by C, with even intermediates suppressed. The two maps are related by $S(n) = C^{v_2(3n+1)+1}(n)$ for all $n \in \mathbb{O}$.

The Terras map $T : \mathbb{Z}^+ \to \mathbb{Z}^+$ *[1] is defined by*

$$T(n) = \begin{cases} n/2 & \text{if } n \text{ is even,} \\ (3n+1)/2 & \text{if } n \text{ is odd.} \end{cases}$$

Note that $T(n)$ is not generally odd when n is odd; the Terras map steps from odd integer to integer (odd or even), whereas S steps from odd integer to odd integer. Note that T maps odd inputs to integers that may themselves be even, so iterating T does not in general recover S. The correct relation between S and C is $S(n) = C^{v_2(3n+1)+1}(n)$ for all $n \in \mathbb{O}$: one application of C to the odd input n produces $3n + 1$, and $v_2(3n + 1)$ further applications of C perform the mandatory halvings.

2 The Mod-24 Universe

2.1 Residue Table

Every odd integer a belongs to one of twelve odd residue classes modulo 24. The table below records $a \bmod c$ for each $c \in \{2, 3, 4, 6, 8, 12, 16, 24\}$ and each representative $a \in \{0, 1, \ldots, 23\}$. Even rows are white; odd rows are coloured by a single global palette in which colour index equals residue value. Because $a \bmod c = a \bmod c'$ for all $c' \geq c$ once c first resolves a, each odd row's colour is constant from the first resolving column rightward, making the coarsening hierarchy visible at a glance.

a	mod2	mod3	mod4	mod6	mod8	mod12	mod16	mod24
0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1
2	0	2	2	2	2	2	2	2
3	1	0	3	3	3	3	3	3
4	0	1	0	4	4	4	4	4
5	1	2	1	5	5	5	5	5
6	0	0	2	0	6	6	6	6
7	1	1	3	1	7	7	7	7
8	0	2	0	2	0	8	8	8
9	1	0	1	3	1	9	9	9
10	0	1	2	4	2	10	10	10
11	1	2	3	5	3	11	11	11
12	0	0	0	0	4	0	12	12
13	1	1	1	1	5	1	13	13
14	0	2	2	2	6	2	14	14
15	1	0	3	3	7	3	15	15
16	0	1	0	4	0	4	0	16
17	1	2	1	5	1	5	1	17
18	0	0	2	0	2	6	2	18
19	1	1	3	1	3	7	3	19
20	0	2	0	2	4	8	4	20
21	1	0	1	3	5	9	5	21
22	0	1	2	4	6	10	6	22
23	1	2	3	5	7	11	7	23

2.2 Mod-24 Adjacency Table

The table below shows, for each residue class $r \in \{0, 1, \ldots, 23\}$, the images of the representative $24a + r$ under C and S. Even residues: $C(24a + r) = 12a + r/2$; S is undefined (input is even). Odd residues: $C(24a + r) = 72a + (3r + 1)$; $S(24a + r)$ divides out all factors of 2. For $r \equiv 5 \pmod 8$ (i.e. $r \in \{5, 13, 21\}$), the value $3r + 1$ is divisible by 8, so $v_2(72a + (3r + 1)) \geq 3$ and the residual v_2 depends on a; the partially-reduced form $(9a+(3r+1)/8)/2^{v_2(9a+(3r+1)/8)}$ is shown for all three cases, with the residual valuation left explicit. In the S mod 8 and S mod 24 columns, r' is a cell-local integer index (a linear function of a) defined to be the unique integer satisfying the displayed parametric form in that cell. For example, for $r = 1$: $S(24a + 1) = 18a + 1$, so the S mod 8 cell uses $r' = 9a$ (giving $2r' + 1 = 18a + 1$), while the S mod 24 cell uses $r' = 3a$ (giving $6r' + 1 = 18a + 1$). The forms $2r' + 1$, $4r' + 1$, $4r' + 3$ indicate the output residue stratum (all odd; $\equiv 1 \pmod 4$; $\equiv 3 \pmod 4$) without specifying which element; r' is the index within that stratum.

r	$r \bmod 8$	$r \bmod 4$	$r \bmod 3$	$24a+r$	$C(24a+r)$	$S(24a+r)$	$S \bmod 8$	$S \bmod 24$
0	0	0	0	$24a$	$12a$	—	—	—
1	1	1	1	$24a+1$	$72a+4$	$18a+1$	$2r'+1$	$6r'+1$
2	2	2	2	$24a+2$	$12a+1$	—	—	—
3	3	3	0	$24a+3$	$72a+10$	$36a+5$	$4r'+1$	$12r'+5$
4	4	0	1	$24a+4$	$12a+2$	—	—	—
5	5	1	2	$24a+5$	$72a+16$	$(9a+2)/2^{v_2(9a+2)}$	$2r'+1$	$2r'+1$
6	6	2	0	$24a+6$	$12a+3$	—	—	—
7	7	3	1	$24a+7$	$72a+22$	$36a+11$	$4r'+3$	$12r'+11$
8	0	0	2	$24a+8$	$12a+4$	—	—	—
9	1	1	0	$24a+9$	$72a+28$	$18a+7$	$2r'+1$	$6r'+1$
10	2	2	1	$24a+10$	$12a+5$	—	—	—
11	3	3	2	$24a+11$	$72a+34$	$36a+17$	$4r'+1$	$12r'+5$
12	4	0	0	$24a+12$	$12a+6$	—	—	—
13	5	1	1	$24a+13$	$72a+40$	$(9a+5)/2^{v_2(9a+5)}$	$2r'+1$	$2r'+1$
14	6	2	2	$24a+14$	$12a+7$	—	—	—
15	7	3	0	$24a+15$	$72a+46$	$36a+23$	$4r'+3$	$12r'+11$
16	0	0	1	$24a+16$	$12a+8$	—	—	—
17	1	1	2	$24a+17$	$72a+52$	$18a+13$	$2r'+1$	$6r'+1$
18	2	2	0	$24a+18$	$12a+9$	—	—	—
19	3	3	1	$24a+19$	$72a+58$	$36a+29$	$4r'+1$	$12r'+5$
20	4	0	2	$24a+20$	$12a+10$	—	—	—
21	5	1	0	$24a+21$	$72a+64$	$(9a+8)/2^{v_2(9a+8)}$	$2r'+1$	$2r'+1$
22	6	2	1	$24a+22$	$12a+11$	—	—	—
23	7	3	2	$24a+23$	$72a+70$	$36a+35$	$4r'+3$	$12r'+11$

The mod-8 and mod-24 residues of $S(24a+r)$ depend on $r \bmod 8$:

- $r \equiv 1 \pmod 8$ ($v_2(3r+1)=2$, $S=18a+\beta$): $S \bmod 8 \equiv 2r'+1 \pmod 8$; $S \bmod 24 \equiv 6r'+1 \pmod{24}$.
- $r \equiv 3 \pmod 8$ ($v_2(3r+1)=1$, $S=36a+\beta$): $S \bmod 8 \equiv 4r'+1 \pmod 8$; $S \bmod 24 \equiv 12r'+5 \pmod{24}$.
- $r \equiv 5 \pmod 8$ ($v_2(3r+1)\geq 3$, coefficient of a reduces to 9): further reduction depends on a; $S \bmod 8$ odd; $S \bmod 24$ ranges over all odd residues.
- $r \equiv 7 \pmod 8$ ($v_2(3r+1)=1$, $S=36a+\beta$): $S \bmod 8 \equiv 4r'+3 \pmod 8$; $S \bmod 24 \equiv 12r'+11 \pmod{24}$.

Note that $r=21\equiv 0 \pmod 3$ is the only $5 \bmod 8$ case that is also $0 \bmod 3$; the others ($r=5\equiv 2$, $r=13\equiv 1$) are non-zero mod 3. Consequently $24a+21$ nodes are leaves (in-degree 0 in G_S), while $24a+5$ and $24a+13$ nodes are hubs with full in-degree.

For $r\equiv 5 \pmod 8$ the image reduces to $S(24a+r)=(9a+\beta)/2^{v_2(9a+\beta)}$ where $\beta\in\{2,5,8\}$. Since $9a+\beta$ has the form $(2j+1)\cdot a+(2j+1)$ (odd coefficient of a, odd constant), $v_2(9a+\beta)$ is unbounded across $a\in\mathbb{N}$: for any k there is an arithmetic progression of a-values with $v_2(9a+\beta)\geq k$. Consequently $S \bmod 8$ and $S \bmod 24$ are unrestricted for these three residue classes; the only invariant is the reduced form $(9a+\beta)/2^{v_2(9a+\beta)}$, which is always odd.

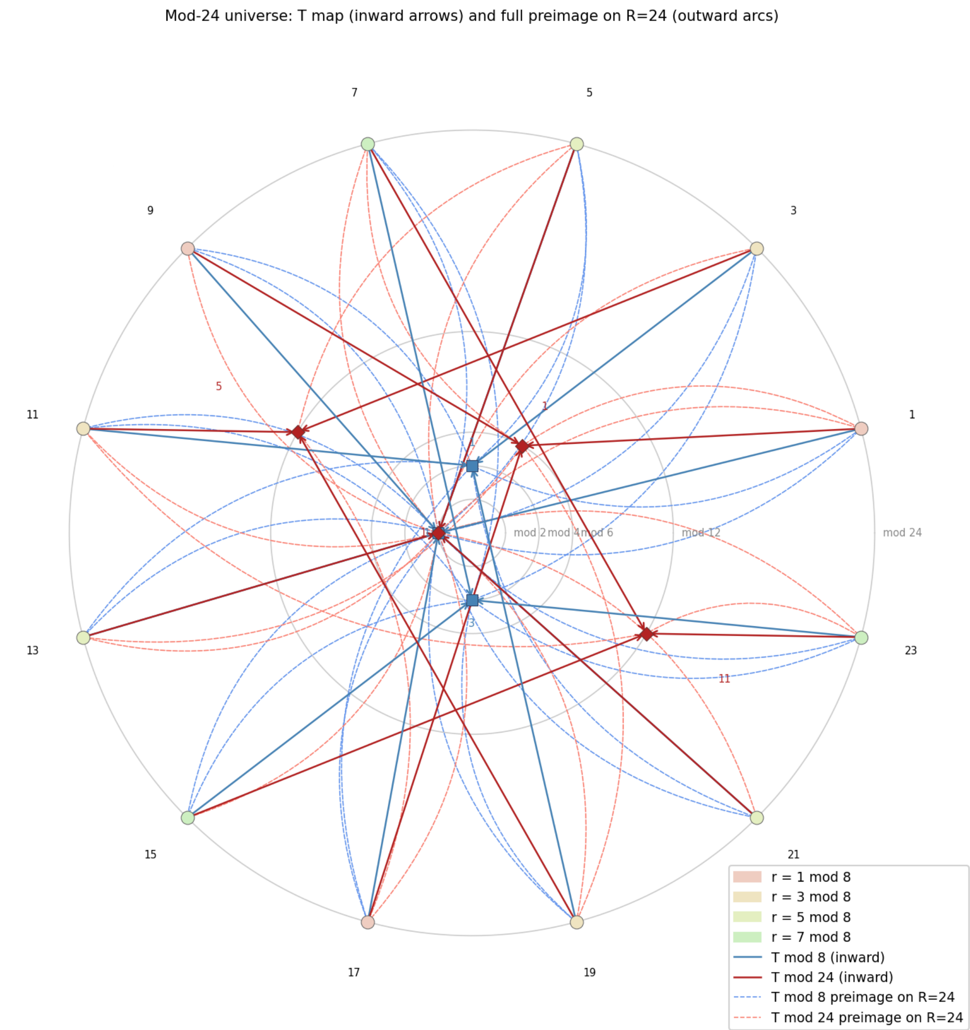

Figure 1: The mod-24 universe. Odd input residues $r \in \{1, 3, \ldots, 23\}$ sit on the outer circle ($R = 24$) at angle $2\pi r/24$. Solid arrows show the image of S: blue to the S mod 8 target on its inner circle, red to the S mod 24 target. Dashed arcs show the full arithmetic-progression preimage on $R = 24$: an output point $kr' + c$ on its circle is connected back to every $v \in \{0, \ldots, 23\}$ with $v \equiv c \pmod{k}$.

2.3 Mod-8 Reduction

Collapsing the mod-24 table to mod 8 by identifying residue classes with the same r mod 8 value yields four odd classes. The S mod 8 column is now exact (not a coarsening), and the mod-24 detail is dropped.

r mod 8	$S(8a + r)$	S mod 8	Note
1	$6a + 1$	$2r'' + 1$	$v_2(3r + 1) = 2$; all odd outputs
3	$12a + 5$	$4r'' + 1$	$v_2(3r + 1) = 1$; outputs $\equiv 1 \pmod{4}$
5	$(3a + 2)/2^{v_2(3a+2)}$	$2r'' + 1$	$v_2(3r + 1) \geq 3$; all odd outputs
7	$12a + 11$	$4r'' + 3$	$v_2(3r + 1) = 1$; outputs $\equiv 3 \pmod{4}$

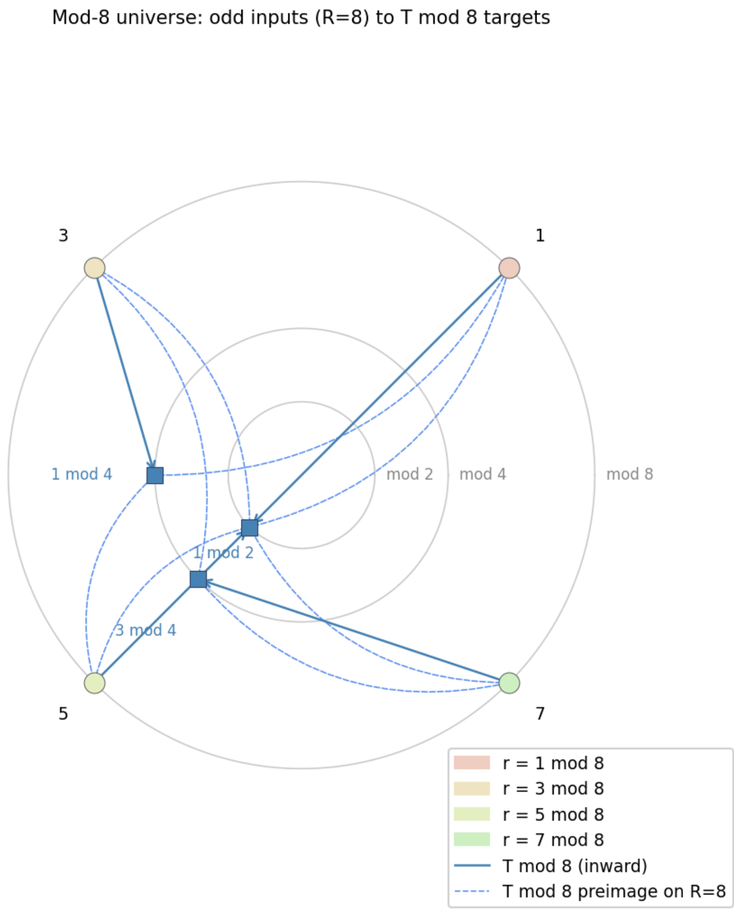

Figure 2: The mod-8 universe. The four odd input residues $\{1, 3, 5, 7\}$ sit on the outer circle $(R = 8)$. Solid arrows show S mod 8 targets on inner circles. Dashed arcs show the full preimage on $R = 8$ of each output class.

2.4 The Mod-8 State Machine

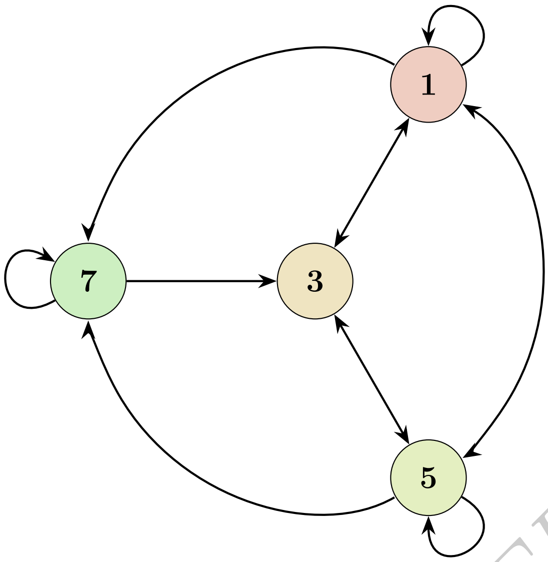

Figure 3: The mod-8 state machine. The forward chain $7 \to 3 \to \{1, 5\}$ funnels low-valuation steps toward the two terminal states, shown stacked in the right column: 1 (top, OEE terminal) and 5 (bottom, OEEE$^+$ terminal). Back-links from 1 (above) and 5 (below) return to any earlier state. Loops on 7, 1, 5: self-transitions are possible. Class 3 has no loop and no back-links: 3-steps always output $\equiv 1 \pmod 4$.

The table below lists every edge of the state machine. The outbound edge taken from any state is selected exactly by

$$S(a) \bmod 8 \;=\; \frac{3a+1}{2^{v_2(3a+1)}} \bmod 8.$$

Remark 2.1 (Threefold symmetry of the mod-8 state machine)**.** *Every structural count in the mod-8 state machine is a multiple of* 3*:*

- ***In-degree 3****: every state has exactly 3 inbound edges (counting self-loops), one from each of its three mod-24 representatives.*
- ***3 self-loops****: states* 7*,* 1*, and* 5 *each have a self-transition; state* 3 *does not.*
- ***3 bidirectional transitions****:* $3 \leftrightarrow 1$*,* $3 \leftrightarrow 5$*, and* $1 \leftrightarrow 5$*.*
- ***3 unidirectional transitions****:* $7 \to 3$*,* $1 \to 7$*, and* $5 \to 7$*.*

State 3 *is the sole exception to self-referentiality: it is a pure funnel, with no self-loop and no outgoing back-links, receiving from* $\{7, 1, 5\}$ *and forwarding only to* $\{1, 5\}$*.*

2.5 Behaviour of the $7 \bmod 8$ Class

Every odd $n \equiv 7 \pmod 8$ satisfies $v_2(3n+1) = 1$, so $S(n) = (3n+1)/2$ and $S(8t+7) = 12t+11$. The output is $\equiv 3 \pmod 8$ when t is even and $\equiv 7 \pmod 8$ when t is odd, so a single step may terminate the run or extend it. The precise mechanism is the content of Theorem 3.6 in Section 3.1.4: the $7 \bmod 8$ run length is the Steiner parameter $\alpha - 2$, governed by the Lyapunov metric $\lambda(n) = v_2(n+1)$.

Corollary 2.2 (Ping-pong exit)**.** *Let* $n = 8t + 7$ *with* $n \equiv 7 \pmod 8$*, and let* $\ell = v_2(n + 1) - 2$ *be the* 7^* *chain length (Theorem 3.6). Then*

$$S^\ell(n) \equiv \begin{cases} 11 \pmod{24} & \textit{if } t \textit{ is even,} \\ 23 \pmod{24} & \textit{if } t \textit{ is odd.} \end{cases}$$

Proof. Each S-step on a 7 mod 8 input maps $t \mapsto \frac{3t+1\cdot\ldots}{2}$ preserving t's parity (since $S(8t+7) = 12t+11$ and $12t+11 = 8(t')+7$ with $t' = \frac{3t+1}{2}$ when t is odd, keeping parity through the chain). The exit value satisfies $S^\ell(n) = 12t' + 11$ for t' of the same parity as t, giving $12t' + 11 \equiv 11$ or $23 \pmod{24}$ as t' is even or odd. □

3 Sequence Encodings of Collatz Paths

A Collatz path from an odd integer a to an odd integer b under the natural map C is a finite sequence of steps, each of which is either an *odd step* ($n \mapsto 3n+1$) or an *even step* ($n \mapsto n/2$). This section develops a hierarchy of notations for such paths, progressing from raw binary strings through run-length encodings to the compressed mod-8 circuit form that underpins the regular expression in Chapter 5.

3.1 Mod-2 Notation

3.1.1 OE Notation

The coarsest encoding records only the parity of each C-step. Write O for an odd step and E for an even step. A path that begins at an odd integer and ends just before the next odd integer is an *OE word*: it consists of exactly one O followed by one or more Es. For example, the path $3 \to 10 \to 5$ is encoded OE, and $3 \to 10 \to 5 \to 16 \to 8 \to 4 \to 2 \to 1$ yields OEEEEE.

A longer orbit is obtained by concatenating OE words. The first few odd iterates of 27 give, in order:

$$27 \xrightarrow{\mathsf{OE}} 41 \xrightarrow{\mathsf{OEE}} 31 \xrightarrow{\mathsf{OEEE}} \cdots$$

Concatenating the words gives the full OE string for the path.

3.1.2 OE Compression

Under Collatz dynamics, every odd step is immediately followed by at least one even step: $3a+1$ is always even, so C must apply at least one halving after every O. The first E after each O is therefore redundant — its presence is guaranteed by the dynamics. *Compressed OE notation* drops this redundant E, replacing each occurrence of OE with a single O and leaving all other symbols unchanged. For example, OEOEE compresses to OOE: each OE becomes O, and the trailing E is retained. The surviving Es are the *excess evens* — the even steps beyond the mandatory minimum — and carry the arithmetic weight of the path.

Compressed OE notation coincides with the notation that arises naturally from the Terras map T [1], in which the odd step is $(3n+1)/2$ rather than $3n+1$: that map folds the mandatory halving into the odd step, so its symbol alphabet records only the excess even halvings. When this paper writes unexponentiated OE strings, however, we always mean the *uncompressed* form — explicit O and E symbols for every individual C-step.

3.1.3 Symbol Exponentiation

Exponential notation operates on the output of compressed OE notation. A run of α consecutive Os is written O^α, and a run of β consecutive Es is written E^β. Continuing the example, OOE becomes $\mathsf{O}^2\mathsf{E}^1$, or simply $\mathsf{O}^2\mathsf{E}$. A Steiner circuit whose compressed form is $\mathsf{O}^\alpha\mathsf{E}^\beta$ has $\alpha = o$ odd steps and $\beta = e - o$ excess even steps (total even count $e = \alpha + \beta$), and satisfies $\beta \geq 1$ (the convergence condition $3^\alpha < 2^{\alpha+\beta}$). The distribution of those β excess even steps among the α odd-step blocks is suppressed in this notation; it is recovered by the mod-8 encoding below.

3.1.4 Steiner Circuits

A *Steiner circuit* is a path whose exponential form is $\mathsf{O}^\alpha\mathsf{E}^\beta$ — that is, all α odd steps precede all β excess even steps in the compressed string. Such paths have a clean arithmetic characterisation that illuminates why the exponential notation is natural.

Proposition 3.1 (Steiner circuit arithmetic)**.** *Let a_0 be a positive odd integer with $a_0+1 = 2^\alpha m$ for some positive odd integer m and $\alpha \geq 1$. Then the orbit of a_0 under C evolves as follows.*

1. *After i OE pairs ($0 \le i \le \alpha$):*

$$a_i + 1 \;=\; 3^i \cdot 2^{\alpha - i} \cdot m.$$

2. *After α OE pairs, $a_\alpha = 3^\alpha m - 1$, which is even.*

3. *The subsequent pure even descent has length $\beta = v_2(3^\alpha m - 1)$, terminating at the odd integer*

$$b \;=\; \frac{3^\alpha m - 1}{2^\beta}.$$

The path from a_0 to b has α odd steps and $e = \alpha + \beta$ even steps in total, and its exponential form is $\mathsf{O}^\alpha \mathsf{E}^\beta$.

Proof. The base case $i = 0$ holds by hypothesis. For the inductive step, suppose $a_i + 1 = 3^i \cdot 2^{\alpha-i} \cdot m$ with $\alpha - i \ge 1$, so a_i is odd (since $3^i \cdot 2^{\alpha-i} \cdot m$ is even). One odd step gives

$$3a_i + 1 \;=\; 3(a_i + 1) - 2 \;=\; 3^{i+1} \cdot 2^{\alpha-i} \cdot m - 2 \;=\; 2\big(3^{i+1} \cdot 2^{\alpha-i-1} \cdot m - 1\big).$$

The mandatory even step divides by 2, giving $a_{i+1} = 3^{i+1} \cdot 2^{\alpha-i-1} \cdot m - 1$, so $a_{i+1} + 1 = 3^{i+1} \cdot 2^{\alpha-(i+1)} \cdot m$, completing the induction.

At $i = \alpha$: $a_\alpha + 1 = 3^\alpha m$, which is odd (both factors odd), so $a_\alpha = 3^\alpha m - 1$ is even and the OE phase ends. The pure even descent lasts exactly $\beta = v_2(3^\alpha m - 1)$ steps by definition of the 2-adic valuation, reaching the odd integer $b = (3^\alpha m - 1)/2^\beta$. The α OE pairs contribute α odd and α even steps; the final descent contributes β more even steps, for $e = \alpha + \beta$ total. In compressed OE notation the mandatory even steps are suppressed, leaving α Os followed by β Es, i.e. $\mathsf{O}^\alpha \mathsf{E}^\beta$. □

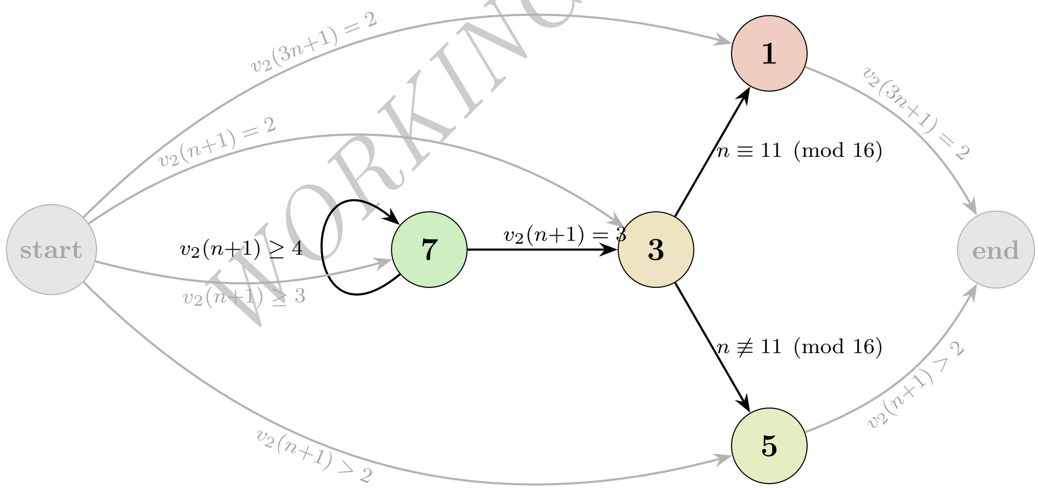

Figure 4: Steiner circuit state machine: the mod-8 machine restricted to paths of the form $(7^*\,3)^?\,(1|5)$. Removed edges: $5 \to 7$, $1 \to 7$, $1 \to 1$, $5 \to 5$, $5 \to 3$, $1 \to 3$, $1 \leftrightarrow 5$. Grey start node connects to every possible circuit entry state; grey end node collects the two circuit exit states 1 and 5.

Remark 3.2 (Steiner circuit machine as a subgraph of the mod-8 machine)**.** *Figure 4 is obtained from the full mod-8 state machine (Figure 3) by deleting seven edges: $5 \to 7$, $1 \to 7$, $1 \to 1$, $5 \to 5$, $5 \to 3$, $1 \to 3$, and $1 \leftrightarrow 5$. The retained edges — the 7 self-loop, $7 \to 3$, $3 \to 1$, and $3 \to 5$ — are exactly those traversed by Steiner circuits. This pattern of restricting the mod-8 machine by edge deletion (and occasionally edge reversal) recurs throughout the paper as a tool for visualising constrained walk families.*

Remark 3.3 (Steiner parameter and the $7 \bmod 8$ run)**.** *The parameter $\alpha = v_2(a_0+1)$ in Proposition 3.1 is the* precise *count of odd steps in the circuit. When every odd step is a $7 \bmod 8$ step (i.e. the circuit is a pure 7^* run followed by a $3 \bmod 8$ exit), the run length is $\alpha - 2$: the circuit uses $\alpha - 2$ steps in the $7 \bmod 8$ class and two further steps to complete the exit transition. This is the content of Theorem 3.6 below.*

Lemma 3.4 (mod-8 class via $v_2(n+1)$)**.** *For any odd integer n:*

$$n \equiv 7 \pmod 8 \iff v_2(n+1) \geq 3, \qquad n \equiv 3 \pmod 8 \iff v_2(n+1) = 2.$$

Proof. Since n is odd, $n+1$ is even. The residue $n \bmod 8 \in \{1,3,5,7\}$ determines $n+1 \bmod 8 \in \{2,4,6,0\}$, and hence $v_2(n+1) \in \{1,2,1,\geq 3\}$ respectively. □

Lemma 3.5 (S decrements $v_2(n+1)$)**.** *If $n \equiv 7 \pmod 8$ then $v_2(S(n)+1) = v_2(n+1) - 1$.*

Proof. Write $n+1 = 2^s q$ with q odd and $s = v_2(n+1) \geq 3$. Since $v_2(3n+1) = 1$ we have $S(n) = (3n+1)/2$, so

$$S(n) + 1 \;=\; \frac{3n+3}{2} \;=\; \frac{3(n+1)}{2} \;=\; 3 \cdot 2^{s-1} q.$$

As $3q$ is odd, $v_2(S(n)+1) = s - 1$. The key identity $S(n) + 1 = \frac{3(n+1)}{2}$ is the fundamental mechanism of every Steiner circuit: each OE pair replaces one factor of 2 in $n+1$ by a factor of 3, transferring the 2-adic weight downward by exactly one. □

Theorem 3.6 (7^* run length and the Steiner parameter)**.** *Let n be an odd integer with $n \equiv 7 \pmod 8$ and $v_2(n+1) = s \geq 3$. Let $\ell(n)$ be the length of the maximal run of S-iterates starting at n that remain $\equiv 7 \pmod 8$:*

$$S^j(n) \equiv 7 \pmod 8 \ (0 \leq j < \ell), \qquad S^\ell(n) \equiv 3 \pmod 8.$$

Then

$$\ell(n) \;=\; v_2(n+1) - 2 \;=\; s - 2.$$

Equivalently, the Steiner parameter of the enclosing circuit satisfies $\alpha = s$, and the 7^* run occupies the first $\alpha - 2$ of the α odd steps.

Proof. By Lemma 3.5, each S-step decrements $v_2(\cdot + 1)$ by 1 while the iterate is in the $7 \bmod 8$ class. After k steps, $v_2(S^k(n)+1) = s - k$. By Lemma 3.4, the iterate remains $\equiv 7 \pmod 8$ iff $s - k \geq 3$, and exits to $\equiv 3 \pmod 8$ when $s - k = 2$, i.e. at $k = s - 2$. □

Remark 3.7 (Parity sequences vs. concrete sequences)**.** *The notation $\mathsf{O}^\alpha \mathsf{E}^\beta$ describes a* parity sequence*: it specifies the shape of the path (how many odd and excess-even steps) but not the concrete starting value. The free parameter is m in the initial condition $a_0 + 1 = 2^\alpha m$. To denote a* concrete *Steiner circuit we write $\mathsf{O}^\alpha \mathsf{E}^\beta(m)$, fixing m and thereby pinning $a_0 = 2^\alpha m - 1$ and all subsequent values uniquely.*

3.2 Mod-8 Notation

3.2.1 Mod-8 Step Alphabet

Moving from the binary alphabet $\{\mathsf{O}, \mathsf{E}\}$ to the four-symbol odd alphabet $\{1,3,5,7\}$ encodes each S-step by the residue class of its input modulo 8. From the mod-8 reduction table (Section 2.3):

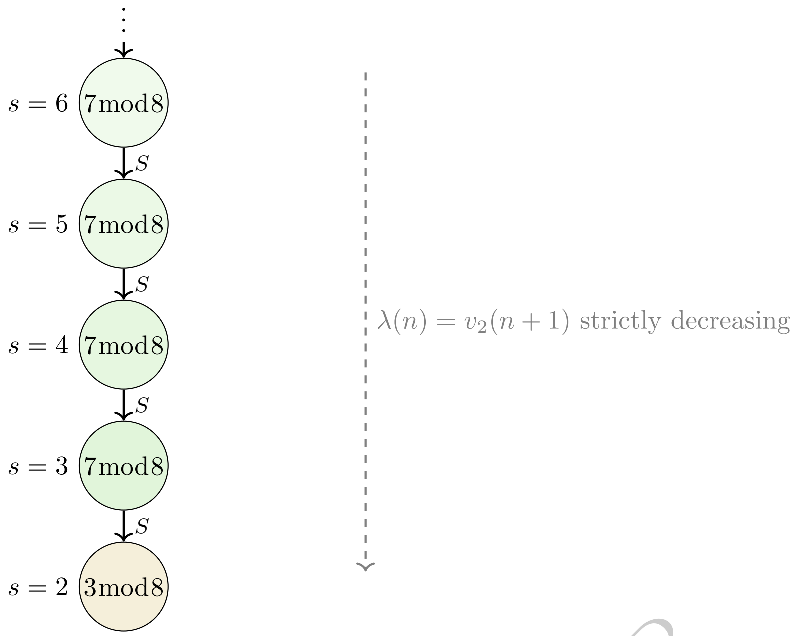

Figure 5: Lyapunov state diagram for 7^* runs (Theorem 3.6). States are labelled by $s = v_2(n+1)$; each S-step decrements s by 1 (Lemma 3.5). The 7 mod 8 class occupies $s \geq 3$; the absorbing state $s = 2$ is the 3 mod 8 exit. Convergence in exactly $s - 2$ steps is guaranteed.

Symbol	Input class	$v_2(3a+1)$
1	$a \equiv 1 \pmod 8$	2
3	$a \equiv 3 \pmod 8$	1
5	$a \equiv 5 \pmod 8$	≥ 3 (varies)
7	$a \equiv 7 \pmod 8$	1

A mod-8 path string is a word over $\{1, 3, 5, 7\}$; for example, `7 3 1` denotes three consecutive S-steps beginning in residue class 7, then 3, then 1.

3.2.2 Compressed Mod-8 Circuit Form

The mod-8 symbols partition into two groups by their contribution to the step budget: *high-excess* steps (5, contributing $v_2 \geq 3$ even steps) and *low-excess* steps $(1, 3, 7$, each contributing $v_2 \leq 2$ even steps).

In a Steiner circuit of parameters (α, β) with $\beta \geq 1$, the circuit terminates with exactly one terminal symbol:

- Symbol 1 (contributing $v_2 = 2$, so two even steps) when $\beta = 1$ (one surplus even step: one of the $v_2 = 2$ drops contributes exactly one net surplus after the S-step accounting).
- Symbol 5 (contributing $v_2 \geq 3$) when $\beta \geq 2$ (the higher valuation absorbs the additional surplus).

The remaining $\alpha - 1$ odd steps are filled from the left by 7-steps (contributing $v_2 = 1$) and a single 3-step (contributing $v_2 = 1$) in the penultimate position (when $\alpha \geq 2$). This yields the *compressed mod-8 circuit form*:

$$7^{\alpha-2}\, 3^{\alpha-1}\, 1^{2-\beta}\, 5^{\beta-1} \tag{1}$$

where $n^k = \varepsilon$ (the empty string) for $k \leq 0$. The three mutual exclusions follow immediately:

β	Expanded	Reading
1	$7^{\alpha-2}\,3^{\alpha-1}\,1^1\,5^0$	$7^{\alpha-2}\,3^{\alpha-1}\,1$
2	$7^{\alpha-2}\,3^{\alpha-1}\,1^0\,5^1$	$7^{\alpha-2}\,3^{\alpha-1}\,5$

Remark 3.8 (Exponent convention)**.** *The convention $n^k = \varepsilon$ for $k \leq 0$ unifies the boolean and arithmetic views: n^{false}, n^0, $n?$ (regex optional), and the empty string ε are all equivalent representations of "zero occurrences of symbol n".*

3.2.3 Connection to the Regular Expression

The compressed form (1) is the bridge to the regular expression characterisation of Chapter 5. Reading off the three syntactic positions:

Position	Compressed	Regex match
Leading 7-steps	$7^{\alpha-2}$	7^* (zero or more)
Penultimate 3-step	$3^{\alpha-1}$	3? (present iff $\alpha \geq 2$)
Terminal symbol	$1^{2-\beta}\,5^{\beta-1}$	(1 | 5) (exactly one)

Each row of the compressed form matches exactly one alternative in the sub-expression $\big((7^*)3?\big)?(1 \mid 5)$, and a full Collatz path is a concatenation of such circuits, giving the language

$$\big((7^*3)?(1 \mid 5)\big)^*.$$

This characterisation is stated as Conjecture 5.1 below; the proof is deferred to [2].

4 Graph-Theoretic Foundations

Readers familiar with directed graph theory will find little that is new in this chapter. The definitions of directed graph, degree, and rooted graph are entirely standard. Two definitions, however, depart from common usage in ways the rest of this paper relies on heavily: the distinction between *path* and *walk* (here paths follow edges and are deterministic; walks oppose them and are not), and the *open/closed* interval notation for paths. Readers comfortable with graph theory are encouraged to skim lightly but to read Definitions 4.4 and 4.5 with care before proceeding.

The default object of study throughout this paper is a *directed graph* (DG), which may contain cycles. A *directed acyclic graph* (DAG) is the important special case in which cycles are absent; its key properties are collected in Definition 4.6 and used where relevant, but they are not assumed to hold by default. The Collatz graph is a DG: it contains at least the cycle arising from the edge $(1, 1)$ in the odd graph (equivalently, $1 \to 4 \to 2 \to 1$ in the natural graph). The Collatz conjecture asserts that deleting this single edge — $(1, 1)$ from G_S, or $(1, 4)$ from G_C — leaves a DAG; see Definition 4.12.

4.1 Directed Graphs: Formal Definitions

We fix notation that will be used throughout this paper. All graphs considered here are *directed*; the qualifier is occasionally omitted when the context is unambiguous.

Definition 4.1 (Directed graph)**.** *A* directed graph *is a pair $G = (V, E)$ where V is a set whose elements are called* nodes *(or* vertices*), and $E \subseteq V \times V$ is a set whose elements are called* directed edges*. An edge $(a, b) \in E$ is said to* leave *a and* enter *b; we write $a \to b$ to denote $(a, b) \in E$. Unless otherwise specified, an edge is represented as the right-open path $[a, b)$: the source node a is included in the edge's node set and the target node b is not.*

Definition 4.2 (In-degree, out-degree, sources and sinks)**.** *The* in-degree *of a node $b \in V$ is the number of edges entering it:*

$$\deg^-(b) \;=\; |\{b' \in V : b' \to b\}|.$$

The out-degree *of a node $a \in V$ is the number of edges leaving it:*

$$\deg^+(a) \;=\; |\{a' \in V : a \to a'\}|.$$

A node a with $\deg^-(a) = 0$ is called an initial node *(or* source*). A node b with $\deg^+(b) = 0$ is called a* terminal node *(or* sink*). The conventions a (source) and b (sink) apply at the level of the graph as a whole. Within local expressions such as edge notation $[a, b)$, a and b denote the left and right endpoints of that edge and need not be global sources or sinks.*

Definition 4.3 (Walk node indexing)**.** *Nodes visited during a walk are written $b_{d,c}$, where $d \geq 0$ is the* depth *of the node relative to the walk's origin and c is a* branch label *that identifies which predecessor was chosen at that depth. The label c is not required to be sequential; it is drawn from an ordered index set whose specific values are determined by context (for example, a Collatz-derived metric may yield labels such as $0, 2, 4, \ldots$ or $1, 3, 5, \ldots$). What is required is only that c uniquely identifies the node among all nodes at depth d reachable from the walk's origin. The walk's origin is b_{0,c_0}; when the origin is the unique sink of a rooted tree, it is written $b_{0,0}$.*

Definition 4.4 (Path and walk)**.** *A* path *of length k in G is a sequence of distinct nodes $a_0, a_1, \ldots, a_k \in V$ such that $a_{i-1} \to a_i$ for each $1 \leq i \leq k$, i.e. a path* follows *the direction of edges. The node $a = a_0$ is the* initial node *and $b = a_k$ is the* terminal node*. Because edges are followed forward, each successor a_i is uniquely determined by a_{i-1} whenever $\deg^+(a_{i-1}) = 1$;*

in a graph where every non-sink node has out-degree 1*, a path is* fully determined *by its initial node* a *and its length* k *(or equivalently by* a *and* b*).*

A walk *of length* k *in* G *is a sequence of distinct nodes* $b_{0,c_0}, b_{1,c_1}, \ldots, b_{k,c_k} \in V$ *such that* $b_{d,c_d} \to b_{d-1,c_{d-1}}$ *for each* $1 \le d \le k$*, i.e. a walk* opposes *the direction of edges, proceeding back toward the sources of the graph. Here* d *denotes the depth of each node relative to the walk's origin* b_{0,c_0}*, and* c_d *is the branch label chosen at depth* d*. Because edges are opposed, each node* b_{d,c_d} *is chosen from* $\{p \in V : p \to b_{d-1,c_{d-1}}\}$*, which may contain more than one node; a walk is therefore* non-deterministic *in general. A walk is fully determined only when a* branch selection procedure *is specified at every node with in-degree greater than* 1*.*

The distinction is definitional and must be maintained throughout: paths follow edges and are deterministic; walks oppose edges and require branch selection to be determined.

Every path $a_0, a_1, \ldots, a_k$ *has a* corresponding walk $b_{0,c_0}, b_{1,c_1}, \ldots, b_{k,c_k}$ *obtained by traversing the same edges in reverse:* $b_{d,c_d} = a_{k-d}$ *for each* $0 \le d \le k$*. The walk is the unique branch-selection of the non-deterministic reverse traversal that recovers exactly the nodes of the path; it is non-deterministic in general because other branch selections at each* b_{d,c_d} *may exist.*

Definition 4.5 (Open and closed paths; interval notation)**.** *Let* $P = (a_0, a_1, \ldots, a_k)$ *be a path with initial node* $a = a_0$ *and terminal node* $b = a_k$*.*

- *A* closed path $[a, b]$ *has node set* $\{a_0, a_1, \ldots, a_k\}$*; both* a *and* b *are included.*
- *A* right-open path $[a, b)$ *has node set* $\{a_0, a_1, \ldots, a_{k-1}\}$*; the path traverses all* k *edges but* b *is excluded from the node set.*
- *A* left-open path $(a, b]$ *has node set* $\{a_1, a_2, \ldots, a_k\}$*; the path traverses all* k *edges but* a *is excluded from the node set.*
- *An* open path (a, b) *has node set* $\{a_1, a_2, \ldots, a_{k-1}\}$*; both endpoints are excluded.*

In every case the word of the path *is the label sequence* $\lambda(a_0 \to a_1)\, \lambda(a_1 \to a_2) \cdots \lambda(a_{k-1} \to a_k)$*, which is the same regardless of which endpoints are included in the node set. A* cycle $[a, a]$ *is a closed path with* $a_0 = a_k = a$*,* $k \ge 1$ *edges, and node set* $\{a_0, a_1, \ldots, a_{k-1}\}$ *as a set (the closing repetition of* a *is not a new node); it follows the direction of edges throughout. Cycles are permitted in a general DG; a graph with no cycles is a DAG (Definition 4.6), in which every path satisfies* $a \ne b$*.*

Definition 4.6 (Directed acyclic graph)**.** *A directed graph* $G = (V, E)$ *is a* directed acyclic graph *(*DAG*) if it contains no cycle, i.e. no closed path* $[a, a]$*. Equivalently, the reachability relation on* V *induced by* E *is a strict partial order. A DAG is a special case of a DG; it gains the additional property that every path reaches a new node at each step.*

Definition 4.7 (Rooted directed graph and the unique-successor property)**.** *A directed graph* $G = (V, E)$ *is* rooted *at* $r \in V$ *if* r *is the unique sink and every node* $a \in V$ *admits at least one path from* a *to* r *(edges point toward* r*, so paths follow them forward to reach* r*). If in addition* $\deg^+(a) = 1$ *for every non-sink node* a*, we say* G *has the* unique-successor property*; the path from any* a *to* r *is then fully determined, and* G *is a* directed tree *rooted at* r*. When the graph is also a DAG, a rooted DG with the unique-successor property is a* rooted directed tree *in the strict sense (no cycles, one path to root).*

Definition 4.8 (Weak connectivity and graph components)**.** *Two nodes* $u, v \in V$ *are* weakly connected *if they are connected by a path or walk when all edges are treated as undirected, i.e. if there exists a sequence* $u = v_0, v_1, \ldots, v_m = v$ *such that for each* $1 \le i \le m$ *either* $v_{i-1} \to v_i$ *or* $v_i \to v_{i-1}$ *holds in* G*. Weak connectivity is an equivalence relation on* V*; its equivalence classes are the* (weakly connected) components *of* G*. A graph is* weakly connected *if it has exactly one component, i.e. every pair of nodes is weakly connected.*

Two nodes $u, v \in V$ are strongly connected *if there exists both a path from u to v and a path from v to u. Strong connectivity is likewise an equivalence relation; its classes are the* strongly connected components *(*SCCs*) of G. In a DAG, every SCC is a single node; non-trivial SCCs (those containing a cycle) can only arise in a general DG.*

4.2 Labelled Graphs and Languages

The graphs arising in this paper carry labels on their edges.

Definition 4.9 (Edge-labelled graph)**.** *Let Σ be a finite set called the* alphabet*. An* edge-labelled directed graph *over Σ is a triple $G = (V, E, \lambda)$ where (V, E) is a directed graph and $\lambda : E \to \Sigma$ assigns a label to each edge.*

Definition 4.10 (Word of a path; language of a node)**.** *Given an edge-labelled directed graph (V, E, λ) and a path $a_0, a_1, \ldots, a_k$, the* word of the path *is the concatenation $\lambda(a_0 \to a_1)\,\lambda(a_1 \to a_2) \cdots \lambda(a_{k-1} \to a_k) \in \Sigma^*$. The* language *of a node a is the set of all words produced by paths starting at a:*

$$\mathcal{L}(a) \;=\; \bigl\{\, s \in \Sigma^* \;\big|\; s \text{ is the word of some path starting at } a \,\bigr\}.$$

4.3 The Collatz Graphs

The maps C and S defined in Definition 1.1 are now instantiated as directed graphs using the framework of the preceding subsections.

Definition 4.11 (Natural and odd Collatz graphs)**.** *The* natural Collatz graph *is the directed graph $G_C = (\mathbb{Z}^+, E_C)$ where $E_C = \{(n, C(n)) : n \in \mathbb{Z}^+\}$.*

The odd Collatz graph *is the directed graph $G_S = (\mathbb{O}, E_S)$ where $E_S = \{(n, S(n)) : n \in \mathbb{O}\}$.*

Both graphs have the unique-successor property: C and S are total maps, so every node has out-degree exactly 1. G_C contains the cycle $1 \to 4 \to 2 \to 1$; G_S inherits this as the edge $(1, S(1)) = (1, 1)$, a self-loop at 1. Both graphs are therefore DGs, not DAGs.

Definition 4.12 (Acyclic restrictions and the Collatz conjecture)**.** *The* acyclic natural graph *is $G'_C = (\mathbb{Z}^+,\ E_C \setminus \{(1, 4)\})$, obtained by deleting the single edge $(1, 4)$ while retaining both nodes. The* acyclic odd graph *is $G'_S = (\mathbb{O},\ E_S \setminus \{(1, 1)\})$, obtained by deleting the self-loop at 1. In both cases 1 becomes a sink and the graph is rooted at 1.*

The Collatz conjecture *asserts that G'_C and G'_S are DAGs, i.e. that $(1, 4)$ is the* only *cycle-forming edge in G_C and that the self-loop at 1 is the only cycle in G_S.*

Definition 4.13 (Collatz paths and walks)**.** *A* natural Collatz path *is a path in G_C; an* odd Collatz path *is a path in G_S. Because both graphs have the unique-successor property, a path in either is fully determined by its initial node and length.*

A natural Collatz walk *is a walk in G_C; an* odd Collatz walk *is a walk in G_S. Both walks are non-deterministic: positive integers may have multiple predecessors under C, and odd integers may have multiple predecessors under S. A walk in either graph is fully determined only when a branch selection procedure is specified at every node with in-degree greater than 1.*

Unless otherwise stated, Collatz path *and* Collatz walk *refer to the odd versions G_S and G'_S.*

The edge labelling that encodes the arithmetic structure of each step of S is introduced in Section 5, where the residue alphabet $\{1, 3, 5, 7\}$ is defined and the main language-theoretic conjecture is stated.

5 Regular Expressions and the Collatz Language

The combinatorial evidence developed in the preceding sections supports the following central conjecture.

Conjecture 5.1 (Regular expression characterisation)**.** *Let $n \in \mathbb{O}$ be any odd positive integer and let $n \to S(n) \to S^2(n) \to \cdots$ be its Syracuse orbit. Encode each iterate by its residue class mod 8 in the alphabet $\{1, 3, 5, 7\}$. Then:*

1. *Every maximal Steiner circuit in the orbit matches the regular expression* $(7^*\,3)?(1 \mid 5)$.
2. *The full sequence of Steiner circuits matches* $\big((7^*\,3)?(1 \mid 5)\big)^*$.

The infrastructure required for a proof — the mod-8 step taxonomy, the Steiner decomposition (Proposition 3.1), the 7-run length theorem (Theorem 3.6), and the mod-24 universe — is fully developed in the preceding sections. The proof itself is deferred to [2].

6 The 5 Mod 8 Overlay Tree

Deferred to [2]. The 5 mod 8 overlay tree is the sub-DAG of the Collatz graph whose nodes are the odd integers $\equiv 5 \pmod 8$ (OEEE$^+$ exit states of Steiner circuits) and whose edges record the branch structure between consecutive such nodes. Its combinatorial properties, including the branch-length distribution, are studied in the companion speculative paper [4].

7 Conclusion

This paper develops the combinatorial infrastructure for a regular-expression description of Collatz orbits.

What is proved.

- The mod-8 step taxonomy (Section 2): the four residue classes $\{1, 3, 5, 7\} \pmod 8$ correspond exactly to the four OE step types.
- The Steiner circuit decomposition (Proposition 3.1): every Collatz orbit decomposes uniquely into Steiner circuits $\mathsf{O}^\alpha\mathsf{E}^\beta$.
- The 7-run length theorem (Theorem 3.6): the length of the maximal $7 \bmod 8$ run in a Steiner circuit is exactly $v_2(n+1) - 2$.
- The mod-24 universe (Section 2): mod-24 is the smallest modulus at which the parity prefix of every odd residue class is fully determined.
- Graph-theoretic foundations (Section 4): definitions and basic properties of the Collatz directed graph G_S.

What is conjectured.

- Conjecture 5.1: every Steiner circuit matches (7*3)?(1|5) and every orbit matches ((7*3)?(1|5))*. Proof deferred to [2].

What is deferred.

- The 5 mod 8 overlay tree and branch structure: deferred to [2] and the speculative paper [4].
- Cycle-theoretic and convergence results: deferred to [3].

References

[1] R. Terras, A stopping time problem on the positive integers, *Acta Arithmetica* **30** (1976), 241–252.

[2] J. Seymour, *A Toolkit for Collatz Path Analysis* (Paper 5), in preparation, 2026.

[3] J. Seymour, *Applications of the mod-8 regular expression language toolbox to the Collatz conjecture* (Paper 8), in preparation, 2026.

[4] J. Seymour, *The* $\frac{1}{3}(3/4)^k$ *Distribution of Steiner Circuit Branch Lengths* (Speculative Results, Paper 65), in preparation, 2026.