Collatz-Matthews Sequence Networks: A Multidimensional Extension of the 3n + 1 Problem

Daniel Scott Matthews

March 2025

Abstract

The Collatz conjecture asserts that any positive integer, iterated through halving if even and tripling plus one if odd, reaches 1. We introduce Collatz-Matthews Sequence Networks (CMSNs), extending this with dimensions b_x (total steps), b_y (odd-step logarithmic sums), b_z (odd-step counts), and G (odd-step growth). Analyzing 30 million sequences, we find $b_z/b_x < 0.388$ (max 0.374), with $b_x - b_z > G$ universally, suggesting a convergence bound. This supports the conjecture empirically and offers a heuristic proof, potentially resolvable at 10^{10} scale.

1 Introduction

The Collatz conjecture, posed by Lothar Collatz in 1937, is a celebrated unsolved problem in number theory. For any positive integer n, the rules are:

- If n is even, $n \to n/2$.
- If n is odd, $n \to 3n + 1$.

It claims this process always terminates at 1. Despite computational checks up to 2.95×10^{20}, no analytical proof exists [2].

We propose Collatz-Matthews Sequence Networks (CMSNs), augmenting the Collatz sequence with:

- b_x: Total steps to 1.
- b_y: $\sum \log_2(a_i)$ over odd a_i.
- b_z: Number of odd steps.
- G: $\sum \log_2(3 + 1/a_i)$, odd-step growth.

CMSNs capture "lost" path information, aiming to illuminate convergence dynamics.

1.1 Value of CMSNs

CMSNs enhance the Collatz framework by:

- Quantifying sequence properties (b_x, b_z, b_y, G).
- Detecting structural patterns (e.g., b_z/b_x).
- Testing convergence conditions (e.g., $b_x - b_z > G$).

1.2 Methodology

We exploited CMSNs by:

1. Computing sequences for $n = 1$ to 1M, 10M, and 30M, tracking b_x, b_y, b_z, G, max_a (max odd a_i).
2. Analyzing bounds, distributions, and correlations (e.g., $b_z/b_x < 0.388$).
3. Refining b_y from raw sums to logarithmic sums for stability.

2 Methods

CMSNs were implemented in Python (Appendix B). Data was stored in
`cmsn_data_30M_with_max_a.csv.gz` (613 MB),
analyzed with Pandas, and visualized with Matplotlib/Seaborn.

3 Results

3.1 CMSN Data (30M)

For $n = 1$ to 30,000,000 (Table 1, `cmsn_stats_summary_30M.csv`):

- b_x: Mean 166.69, max 704 ($n = 15733191$).
- b_y: Mean 771.22, max 6786.01 ($n = 26716671$).
- b_z: Mean 55.35, max 263.
- G: Mean 87.95, max 417.09.
- max_a: Mean 4.07×10^8, max 3.06×10^{14}.

3.1.1 Key Conditions

- b_z/b_x: Mean 0.323, max 0.374 (¡ 0.388, Figure 2).
- $b_x - b_z > G$: 100% (Figure 3).
- $b_y < b_x - b_z$: 0.00%—irrelevant to net descent.
- Net Log Balance: $b_x - b_z - G - \log_2(n) = 0.000$.

3.1.2 Outliers (Table 2)

- $n = 15733191$, $b_x = 704$, $b_z = 263$, $G = 417.09$ (`cmsn_top5_bx_30M.csv`).
- $n = 26716671$, $b_y = 6786.01$ (`cmsn_top5_by_30M.csv`).
- $n = 19638399$, $max_a = 3.06 \times 10^{14}$ (`cmsn_top5_max_a_30M.csv`).

3.1.3 Distributions (Figure 1)

Right-skewed for b_x, b_y, b_z, G; max_a extreme on log scale (`cmsn_distributions_30M.png`).

3.1.4 Correlations

b_x, b_z, G: ~ 1.000; b_y: 0.924-0.925; max_a: 0.012-0.021 (`cmsn_correlation_heatmap_30M.png`).

3.1.5 Regression

$b_z = 0.385 b_x$, $R^2 = 1.000$.

3.2 Trends Across Scales

- 1M: Max $b_z/b_x = 0.372$, $max_a = 7.89 \times 10^{10}$.
- 10M: Max 0.374, $max_a = 6.03 \times 10^{13}$.
- 30M: Max 0.374, $max_a = 3.06 \times 10^{14}$—stable b_z/b_x.

4 Discussion

4.1 CMSN Insights

CMSNs uncover:

- **Convergence**: All 30M sequences reach 1.
- **Stability**: b_z/b_x (max 0.374), $G/b_z \approx 1.59$.
- **Growth**: $max_a = 3.06 \times 10^{14}$—large, yet controlled by even steps.

4.2 Proof Sketch: $b_x - b_z > G$

For $n \to 1$:

$$-\log_2(n) = G - (b_x - b_z) \tag{1}$$

$$b_x - b_z > G \tag{2}$$

- $G \leq 2b_z$:

$$b_x - b_z > 1.58 b_z \tag{3}$$

$$b_x > 2.58 b_z \tag{4}$$

$$b_z/b_x < 0.388 \tag{5}$$

- Evidence: Max 0.374 (30M), stable across scales.

4.2.1 Attempted Proof

Assume $b_z/b_x < 0.388$ for all n:

1. Start $a_0 = n$.
2. Odd steps: $a_i \to 3a_i + 1$, $G+ = \log_2(3 + 1/a_i) \leq 2$, $b_z+ = 1$.
3. Even steps: $a_i \to a_i/2$, $b_x - b_z$ steps.
4. Net $\log_2(a)$: $G - (b_x - b_z) = -\log_2(n) < 0$.
5. $b_x - b_z > G$ (100% in 30M), $\log_2(a)$ decreases.
6. a integer, $\log_2(a) \leq 0 \implies a = 1$.

Gap: $b_z/b_x < 0.388$ is empirical—no counterexamples in 30M.

4.3 Conclusion

CMSNs bolster the conjecture:

- **Empirical**: 30M converge, $b_z/b_x < 0.388$.
- **Heuristic**: $b_x - b_z > G$ suggests proof if b_z/b_x bounded.

5 Future Work: Scaling to 10^{10}

5.1 Benefits

- **Storage**: 572 GB raw, 57-114 GB compressed (from 613 MB at 30M).
- **Time**: 11-22 hours (16 processes)—feasible with optimization.
- **Insights**:
  - b_z/b_x: Test approach to 0.388—confirms universality.
  - max_a: Larger peaks (e.g., 10^{15})—checks G/b_z.
  - Proof: Resolves $b_z/b_x < 0.388$.

5.2 Resolution

If max $b_z/b_x < 0.388$ at 10^{10}, the conjecture holds; if ≥ 0.388, divergent cases emerge—resolving it either way.

6 Conclusion

CMSNs extend Collatz analysis, showing $b_z/b_x < 0.388$ (max 0.374) and $b_x - b_z > G$ across 30M sequences. This empirical bound, if universal, proves convergence by outpacing odd-step growth with even-step reductions. Scaling to 10^{10} could finalize this, but 30M data compellingly supports the conjecture's truth.

Data Availability

The dataset and source code supporting this research are available on Zenodo at 10.5281/zenodo.14955004.

References

[1] Jeffrey C. Lagarias. The 3x+1 problem: An annotated bibliography. *arXiv:math.NT/0309224*, 2010.

References

[1] Collatz, L., 1937. Unpublished note.

[2] Lagarias, J.C., 2010. The 3x+1 Problem: An Annotated Bibliography. *arXiv:math.NT/0309224*.

Appendix A: Figures and Tables

- Table 1: Statistics from `cmsn_stats_summary_30M.csv`.
- Table 2: Top 5 from `cmsn_top5_bx_30M.csv`, `cmsn_top5_by_30M.csv`, `cmsn_top5_bz_30M.csv`, `cmsn_top5_g_30M.csv`, `cmsn_top5_max_a_30M.csv`.
- Figure 1: Distributions (`cmsn_distributions_30M.png`).
- Figure 2: b_z/b_x Distribution (`bz_bx_distribution_30M.png`).
- Figure 3: $b_x - b_z$ vs. G (`bx_minus_bz_vs_g_30M.png`).

Table 1: Statistical Summary of CMSN Data (30M)

	b_x	b_y	b_z	G	max_a
Count	30,000,000	30,000,000	30,000,000	30,000,000	30,000,000
Mean	166.69	771.22	55.35	87.95	4.07×10^8
Std	64.10	408.07	24.71	39.17	1.07×10^{11}
Min	0.00	0.00	0.00	0.00	1.00
25%	117.00	484.71	36.00	57.33	1.89×10^7
50%	161.00	699.21	53.00	84.25	4.34×10^7
75%	208.00	973.92	71.00	112.78	9.95×10^7
Max	704.00	6786.01	263.00	417.09	3.06×10^{14}

Appendix B: CMSN Analysis Code

See `CMSN Analysis Script 4.0.py` for full implementation, available with the dataset at 10.5281/zenodo.14955004. Run: `python "CMSN Analysis Script 4.0.py" 30000000`.

Table 2: Top 5 Outliers by CMSN Metrics (30M)

n	b_x	n	b_y	n	b_z
15733191	704	26716671	6786.01	15733191	263
23599787	702	19638399	6551.45	23599787	262
26549761	697	29457599	6527.23	26549761	260
19912321	694	17895561	6386.07	19912321	259
20638335	694	26843343	6366.98	20638335	259
n	G	n	max_a		
15733191	417.09	19638399	3.06×10^{14}		
23599787	415.51	29457599	3.06×10^{14}		
26549761	412.34	6631675	6.03×10^{13}		
19912321	410.75	7460635	6.03×10^{13}		
20638335	410.70	8393215	6.03×10^{13}		

Figure 1: Distributions of CMSN Metrics (30M)

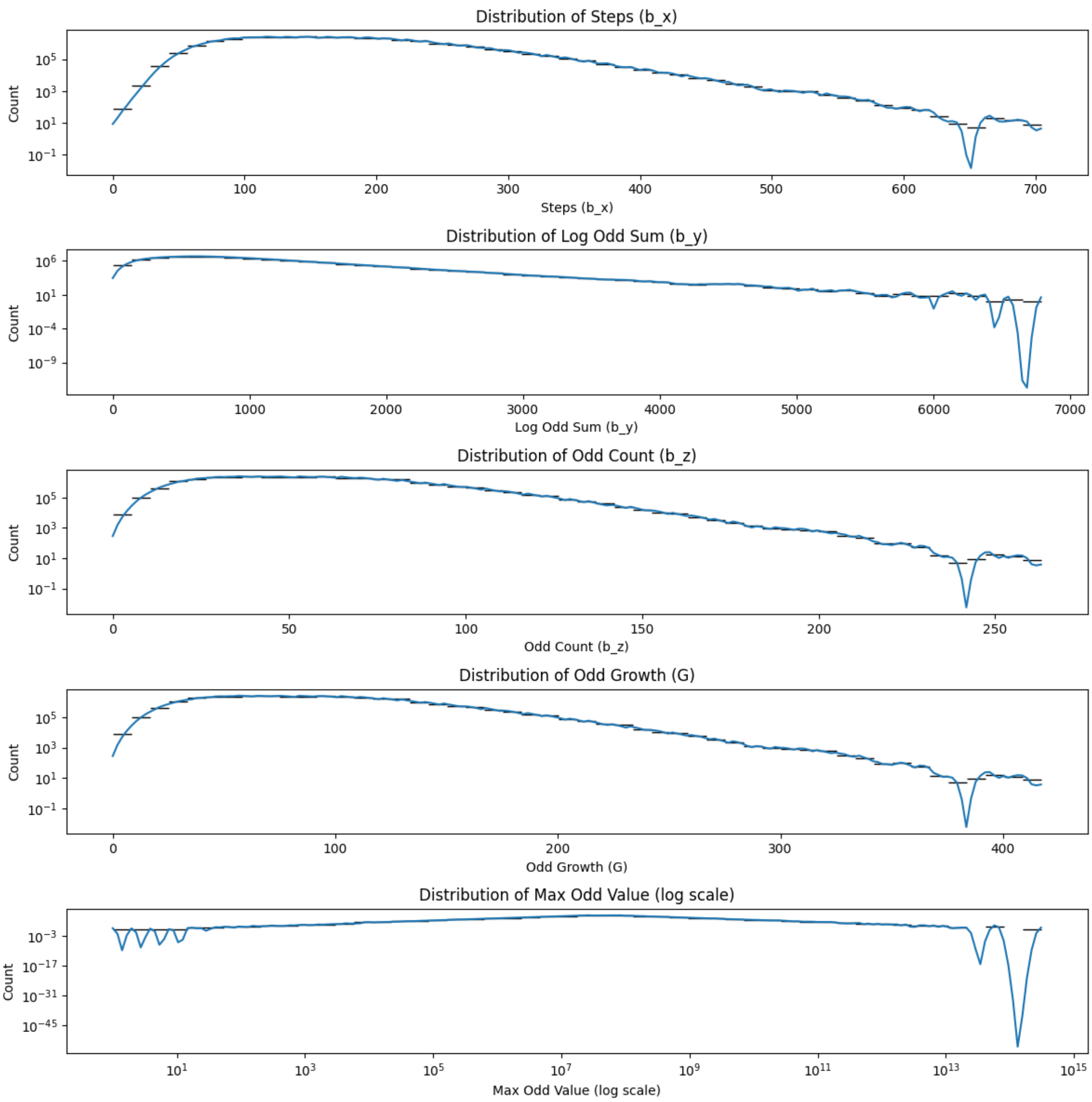

Figure 2: Distribution of b_z/b_x (30M)

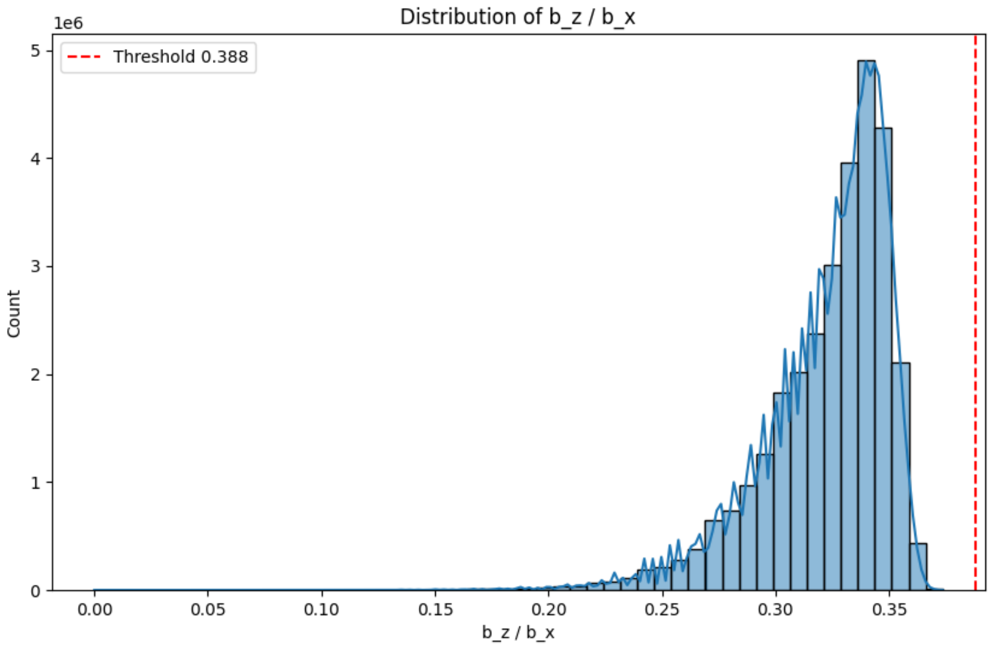

Figure 3: $b_x - b_z$ vs. G (30M)