Head Bites Tail:

Crunching Numbers to Match Integer Sequences

Hieu Nguyen

Rowan University

Math Department Colloquium

December 8, 2011

ABSTRACT: In this talk I will discuss an algorithm to experimentally match integer sequences as part of an ongoing project to mine the Online Encyclopedia of Integer Sequences for new identities. In particular, a similarity measure called *head-bites-tail overlap* will be introduced and shown how to compute distance between two finite sequences and calculate a match probability. Examples of some experimental conjectures found using a *Mathematica* implementation of this algorithm will be presented. This talk is highly accessible to students: those having a background in high school algebra should be able to understand most of this talk and those with a background in discrete math and introductory computer programming should fully appreciate this talk.

Online Encyclopedia of Integer Sequences (OEIS)

- **Searchable online database - http://oeis.org/**
- **Contains almost 200,000 integer sequences**
- **Created by Neil Sloane (AT & T Bell Labs)**
- **Maintained by OEIS Foundation**
- **Example: F_n = 0, 1, 1, 2, 3, 5, 8, 13, 21, ...**

Mining the OEIS

- **Data Mining (Large Scale Pattern Recognition)**

  Process of extracting patterns from large datasets using computer science, mathematics, and statistics.

- **Mine OEIS for Integer Sequence Identities**
- **Enlarge OEIS database to include sequence transformations**
- **Find matches between integer sequences (experimental conjectures)**
- **Prove experimental conjectures that are interesting to obtain new identities**
- **GOAL: Discover interesting connections between different areas of mathematics**

Experimental Pattern Matching

- **Example 1**

- **A000045 : Fibonacci sequence**
  $F_n = 0, 1, 1, 2, 3, 5, 8, 13, 21, \ldots, 39088169$ (39 terms); $n \geq 0$

  A000045S1T3: Sums of Squares Transformation
  $\sum_{k=0}^{n} F_k^2 = 0, 1, 2, 6, 15, 40, 104, \ldots, 2472169789339634$; $n \geq 0$

  A000045S1T8: Product of Consecutive Terms Transformation
  $F_n F_{n+1} = 0, 1, 2, 6, 15, 40, 104, \ldots, 2472169789339634$; $n \geq 0$

  EXPERIMENTAL CONJECTURE: $\boxed{\sum_{k=0}^{n} F_k^2 = F_n \, F_{n+1}}$

- **Example 2**

- **A131524: Number of possible palindromic rows in an n X n crossword puzzle**
  a_n = 0, 0, 1, 1, 2, 2, 4, 4, 7, 7, 12, ..., 121392; $n \geq 1$ (50 terms)

  A131524S2T4: Binomial Transform of a_{2n} (pad $a_0 = 0$):

  $\sum_{k=0}^{n} (-1)^k \binom{n}{k} a_{2k} = 0, 0, 1, 1, 2, 3, 5, 8, 13, \ldots 4181; n \geq 0$

- **A018910S1T4: Pisot sequence L(4,5)**
  b_n = 4, 5, 7, 10, 15, 23, 36, 57, ..., 165580143 $n \geq 0$ (39 terms)

  A018910S1T4: Binomial Transform of b_n:

  $\sum_{k=0}^{n} (-1)^k \binom{n}{k} b_k = 4, -1, 1, 0, 1, 1, 2, 3, 5, 8, 13, \ldots, 4181, \ldots, 9227465 \quad (n \geq 0)$

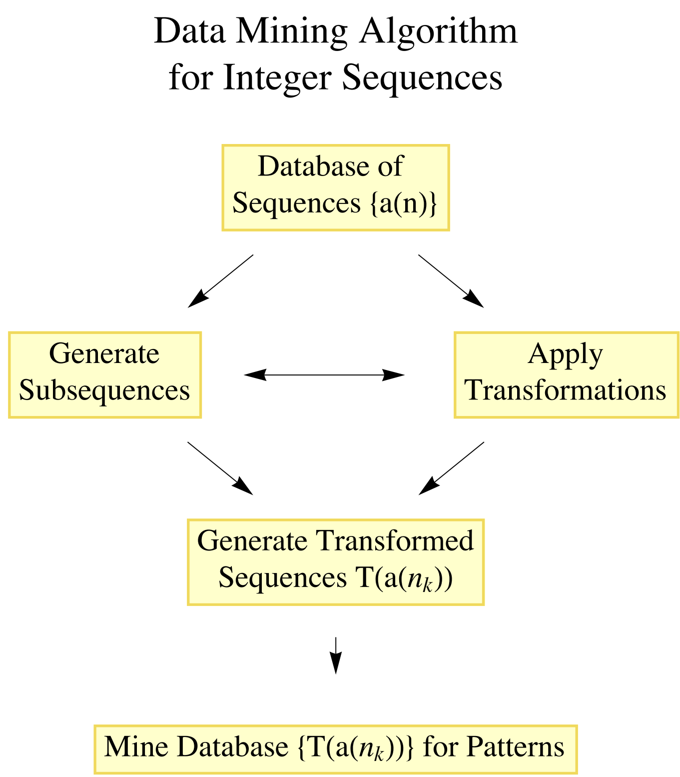
  EXPERIMENTAL CONJECTURE: $\boxed{\sum_{k=0}^{n} (-1)^n \binom{n}{k} a_{2k} = F_{n-1} = \sum_{k=0}^{n+2} (-1)^{n+2} \binom{n+2}{k} b_k} \quad (n \geq 1)$

Hunting for Identities

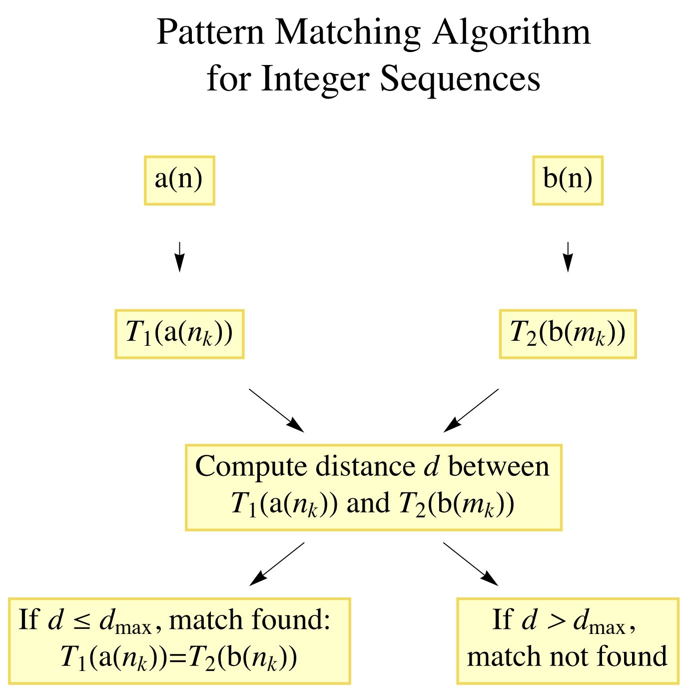
- **Classical Approach**
- **Modern Approach**

  Small-scale (human) versus large-scale (computer)

Out[108]=



Out[129]=


Pattern Matching Algorithm
for Integer Sequences
a(n)
b(n)
$T_1(a(n_k))$
$T_2(b(m_k))$
Compute distance d between $T_1(a(n_k))$ and $T_2(b(m_k))$
If $d \le d_{max}$, match found: $T_1(a(n_k))=T_2(b(n_k))$
If $d > d_{max}$, match not found

Database of Sequence Transformations

- **Source Data - OEIS**
- **Set of Transformations**

LABEL	TRANSFORMATION	FORMULA
T1	Identity	$a(n)$
T2	Partial Sums	$\sum_{k=0}^{n} a(k)$
T3	Partial Sums of Squares	$\sum_{k=0}^{n} a(k)^2$
T4	Binomial Transform	$\sum_{k=0}^{n} (-1)^k \binom{n}{k} a(k)$
T5	Self – Convolution	$\sum_{k=0}^{n} a(k)\, a(n-k)$
T6	Linear Weighted Partial Sums	$\sum_{k=1}^{n} k\, a(k)$
T7	Binomial Weighted Partial Sums	$\sum_{k=0}^{n} \binom{n}{k} a(k)$
T8	Product of Consecutive Elements	$a(n)\, a(n+1)$
T9	Cassini	$a(n-1)\, a(n+1) - a(n)^2$
T10	First Stirling	$\sum_{k=0}^{n} s(n, k)\, a(k)$
T11	Second Stirling	$\sum_{k=0}^{n} S(n, k)\, a(k)$

- **Create MySQL Database of Sequence Transformations**

Acknowledgement: Doug Taggart (Undergraduate Research Assistant)

ID	Label	Subsequence	Transformation	Position	Entry1	Entry2	Entry3
1	A000045S1T1	1	1	0	0	1	1
2	A000045S1T1	1	1	1	1	1	2
3	A000045S1T1	1	1	2	1	2	3
4	A000045S1T1	1	1	3	2	3	5
5	A000045S1T1	1	1	4	3	5	8
...	...	...	...	...	...	...	...
38	A000045S1T1	1	1	37	24 157 817	39 088 169	Null
39	A000045S1T1	1	1	38	39 088 169	Null	Null

Matching Integer Sequences

- **Exercise:**

Consider the finite sequence $\{a(n)\} = \{1, 1, 2, 3, 5, 8, 13, 21, 34, 55\}$. Compare $a(n)$ with each of the four finite sequences below, which are similar to $a(n)$ but do not match exactly. Is there a way to measure how close each sequence matches with $a(n)$ in the sense that both are likely to be subsets of the same infinite sequence (namely the Fibonacci sequence)? If so, then which sequence matches *best* with $a(n)$?

1. $\{1, 1, 2, 3, 5, 8, 13, 21, 47, 55\}$
2. $\{55, 89, 144, 233, 377, 610\}$
3. $\{3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377\}$
4. $\{2, 3, 5, 8, 13, 21, 34\}$
5. $\{1, 0, 1, 1, 2, 3, 5, 8, 13\}$

- **Mathematical Model:**

Determine an appropriate *distance function* (or *similarity measure*) to match two sequences that are *similar*, but not exactly the same.

Overlap

- **Main Assumption:**

Perfect data set - no errors in the values of each integer sequence

- **Overlapping Run**

1. $\{\mathbf{1, 1, 2, 3, 5, 8, 13, 21}, 47, \mathbf{55}\}$
$\{a(n)\} = \{\mathbf{1, 1, 2, 3, 5, 8, 13, 21}, 34, \mathbf{55}\}$
NO MATCH (Worst)

2. $\{\mathbf{55}, 89, 144, 233, 377, 610\}$
$\{a(n)\} = \{1, 1, 2, 3, 5, 8, 13, 21, 34, \mathbf{55}\}$
MATCH

3. $\{\mathbf{3, 5, 8, 13, 21, 34, 55}, 89, 144, 233, 377\}$
$\{a(n)\} = \{1, 1, 2, \mathbf{3, 5, 8, 13, 21, 34, 55}\}$
MATCH

4. $\{\mathbf{2, 3, 5, 8, 13, 21, 34}\}$
$\{a(n)\} = \{1, 1, \mathbf{2, 3, 5, 8, 13, 21, 34}, 55\}$
MATCH (Best)

Ouroboros (Bites Tail)

- **What qualifies as a match between two finite sequences?**

$$\left\{\overset{\text{Head}}{a(1)},\ a(2),\ \ldots,\ a(N-1),\ \overset{\text{Tail}}{a(N)}\right\}$$

$$\left\{\underset{\text{Head}}{b(1)},\ b(2),\ \ldots,\ b(M-1),\ \underset{\text{Tail}}{b(M)}\right\}$$

We will say that two sequences *likely match* or are *similar* (in the sense that there is a chance that both finite sequences are part of the same infinite sequence) if the **head** (beginning) of one sequence **bites** (overlaps with) the **tail** (end) of the other sequence.

Head-Bites-Tail Overlap

INFORMAL DEFINITION: We say that two finite sequences contain a *head-bites-tail (HBT) overlap* if there is an overlapping run which starts at the beginning of one sequence and stops at the end of either sequence.

Let L denote the length of an HBT overlap. There are four cases to consider:

CASE 1a: $L = N - n_0 + 1$

$$\begin{array}{ll} \mathtt{a(1),a(2),...} & \boxed{\mathtt{a(n_0),...,a(N)}} \\ & \boxed{\mathtt{b(1),...,\ b(L)}}\mathtt{,...b(M)} \end{array}$$

CASE 1b: $L = M$

$$\begin{array}{ll} \mathtt{a(1),a(2),...} & \boxed{\mathtt{a(n_0),...,a(n_0+M-1)}}\mathtt{,...,a(N)} \\ & \boxed{\mathtt{b(1),...,\ b(M)}} \end{array}$$

CASE 2a: $L = M - m_0 + 1$

$$\begin{array}{ll} & \boxed{\mathtt{a(1),...,\ a(L)}}\mathtt{,...a(N)} \\ \mathtt{b(1),b(2),...} & \boxed{\mathtt{b(m_0),...,b(M)}} \end{array}$$

CASE 2b: $L = N$

$$\begin{array}{ll} & \boxed{\mathtt{a(1),...,\ a(N)}} \\ \mathtt{b(1),b(2),...} & \boxed{\mathtt{b(m_0),...,b(m_0+N-1)}}\mathtt{,...,b(M)} \end{array}$$

Maximum HBT Overlap

Let $\{a(n)\}_{n=1}^{N}$ and $\{b(m)\}_{m=1}^{M}$ be two finite sequences.

DEFINITION: We say that $a(n)$ and $b(n)$ contain a *head-bites-tail (HBT) overlap* of length L if one of the following two conditions hold:

1. $a(N-L+k)=b(k)$ for all $k=1, \ldots, L$ or $a(n_0+k-1)=b(k)$ for a fixed positive integer n_0 and all $k=1, \ldots, L$.

2. $a(k)=b(M-L+k)$ for all $k=1, \ldots, L$ or $a(k)=b(m_0+k-1)$ for a fixed positive integer m_0 and all $k=1, \ldots, L$.

DEFINITION: We define $L_{\max}$ to be the *maximum HBT overlap*, i.e. the length of the longest HBT overlap, between $a(n)$ and $b(n)$. If no HTB overlap exists, then we set $L_{\max}=0$.

Examples

1. $\{a(n)\}=\{1, 1, 2, 3, 5, 2, 3, 5, \mathbf{2, 3, 5}\}$
   $\{b(n)\}=\qquad\qquad\qquad\qquad\{\mathbf{2, 3, 5}, 2, 3, 5\}$
$L=3$

2. $\{a(n)\}=\{1, 1, 2, 3, 5, \mathbf{2, 3, 5, 2, 3, 5}\}$
   $\{b(n)\}=\qquad\qquad\qquad\{\mathbf{2, 3, 5, 2, 3, 5}\}$
$L_{\max}=6$

HBT Distance

DEFINITION: We define the *head-bites-tail (HBT) distance d* between $a(n)$ and $b(n)$ to be

$$d := d(a(n), b(n)) = N + M - 2\,L_{\max}$$

where $L_{\max}$ is the maximum HBT overlap between $a(n)$ and $b(n)$.

NOTE: d can also be thought of as specifying the number of remaining elements in $a(n)$ and $b(n)$ that DO NOT overlap.

- **Examples**

1. $\{a(n)\} = \{\mathbf{55}, 89, 144, 233, 377, 610\}$
   $\{b(n)\} = \{1, 1, 2, 3, 5, 8, 13, 21, 34, \mathbf{55}\}$
   $d = 6 + 10 - 2\,(1) = 14$

2. $\{a(n)\} = \{\mathbf{3, 5, 8, 13, 21, 34, 55}, 89, 144, 233, 377\}$
   $\{b(n)\} = \{1, 1, 2, \mathbf{3, 5, 8, 13, 21, 34, 55}\}$
   $d = 11 + 10 - 2\,(7) = 7$

3. $\{a(n)\} = \{\mathbf{2, 3, 5, 8, 13, 21}\}$
   $\{b(n)\} = \{1, 1, \mathbf{2, 3, 5, 8, 13, 21}, 55, 81\}$
   $d = 6 + 10 - 2\,(6) = 4$

4. $\{a(n)\} = \{\mathbf{2, 3}\}$
   $\{b(n)\} = \{1, 1, \mathbf{2, 3}, 5, 8\}$
   $d = 2 + 6 - 2\,(2) = 4$

Relative HBT Distance

DEFINITION: We define the *relative HBT distance* r between $a(n)$ and $b(n)$ to be

$$d_r := r(a(n), b(n)) = \frac{d}{N+M} = \frac{N+M-2L}{N+M} = 1 - \frac{2L}{N+M}$$

NOTE: $0 \leq r \leq 1$

DEFINITION: We define the *HBT probability of match* p between $a(n)$ and $b(n)$ to be

$$p := p(a(n), b(n)) = 1 - r = \frac{2L}{N+M}$$

- **Examples**

1. $\{a(n)\} = \{\mathbf{55}, 89, 144, 233, 377, 610\}$
   $\{b(n)\} = \{1, 1, 2, 3, 5, 8, 13, 21, 34, \mathbf{55}\}$
   $d_r = \frac{6+10-2(1)}{6+10} = \frac{14}{16} = \frac{7}{8}$
2. $\{a(n)\} = \{\mathbf{3, 5, 8, 13, 21, 34, 55}, 89, 144, 233, 377\}$
   $\{b(n)\} = \{1, 1, 2, \mathbf{3, 5, 8, 13, 21, 34, 55}\}$
   $d_r = \frac{11+10-2(7)}{11+10} = \frac{7}{21} = \frac{1}{3}$
3. $\{a(n)\} = \{\mathbf{2, 3, 5, 8, 13, 21}\}$
   $\{b(n)\} = \{1, 1, \mathbf{2, 3, 5, 8, 13, 21}, 55, 81\}$
   $d_r = \frac{6+10-2(6)}{6+10} = \frac{4}{16} = \frac{1}{4}$
4. $\{a(n)\} = \{\mathbf{2, 3}\}$
   $\{b(n)\} = \{1, 1, \mathbf{2, 3}, 5, 8\}$
   $d_r = \frac{2+6-2(2)}{2+6} = \frac{4}{8} = \frac{1}{2}$

HBT Conjecture

HBT CONJECTURE: $d(\cdot,\ \cdot)$ is a distance function, i.e. d satisfies the three properties:
I. Positive-definiteness: $d(a(n),\ b(n)) \geq 0$ and $d(a(n),\ b(n)) = 0$ iff $a(n) = b(n)$
II. Symmetry: $d(a(n),\ b(n)) = d(b(n),\ a(n))$
III. Triangle inequality: $d(a(n),\ b(n)) \leq d(a(n),\ c(n)) + d(c(n),\ b(n))$

NOTE: Evidence suggests that HBT Conjecture is true for the space of monotone sequences.

- **Example: Triangle Inequality**

$\{a(n)\} = \{1, 1, 2, 3, 5, 8, 13, 21, 47, 55\}$
$\{b(n)\} = \qquad\qquad\qquad\qquad \{55, 89, 144, 233, 377, 610\}$
$\{c(n)\} = \qquad \{3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377\}$

$$d = N + M - 2\, L_{\max}$$

$d(a(n),\ b(n)) = 10 + 6 - 2\,(1) = 14 = 9_{\text{left}} + 5_{\text{right}}$

$d(a(n),\ c(n)) = 10 + 11 - 2\,(7) = 7 = 3_{\text{left}} + 4_{\text{right}}$

$d(c(n),\ b(n)) = 11 + 6 - 2\,(5) = 7 = 6_{\text{left}} + 1_{\text{right}}$

$$\therefore\ d(a(n),\ b(n)) \leq d(a(n),\ c(n)) + d(c(n),\ b(n))$$

Mathematica Implementation of Maximum HBT Distance

Algorithm for finding $L_{\max}$ (maximum HBT distance)

$\{u(n)\}_{n=1}^{N} = \{1, 1, 2, 3, 5, 8, 13, 21, 47, 55\}$

$\{v(m)\}_{m=1}^{M} = \{3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377\}$

1. Take last element $u(N)$ and find its occurrences in $\{v(m)\}$. Denote the positions of these occurrences by $\{p_k\}_{k=1}^{K}$ (decreasing order).

2. Loop through $k = 1, \ldots, K$:
If $\{u(N - p_k + 1), u(N - p_k + 2), \ldots, u(N)\} = \{v(1), v(2), \ldots, v(p_k)\}$, then $u(n)$ and $v(n)$ have an HBT overlap of length p_k.

3. Repeat steps 1 and 2, but switch roles of $u(n)$ and $v(n)$.

4. Set $L_{\max}$ equal to the length of the longest HBT overlap obtained from steps 1-3.

Mathematica Module

In[131]:=
```
Clear[HBTdistance];
HBTdistance[u_,v_]:=Module[{lengthu,lengthv,positionlastuinv,positionlastvinu,
match,distance,rdistance,i,p,overlap1,overlap2,overlaptemp},

lengthu=Length[u];
lengthv=Length[v];
positionlastuinv=Flatten[Position[v,u[[lengthu]]]];
positionlastvinu=Flatten[Position[u,v[[lengthv]]]];
Print["N = ",lengthu," ; ","M = ",lengthv];

match=0;
overlap1=0;
If[positionlastuinv!={},
    i=1;
    While[match==0&&i<=Length[positionlastuinv],
        p=positionlastuinv[[-i]];
        overlaptemp=Min[lengthu,p];
        If[Take[u,-overlaptemp]==Take[v,{p-overlaptemp+1,p}],
            match=1;overlap1=overlaptemp,
            i++
        ]
    ]
];

match=0;
overlap2=0;
If[positionlastvinu!={},
    i=1;
```

```
        While[match==0&&i<=Length[positionlastvinu],
            p=positionlastvinu[[-i]];
            overlaptemp=Min[lengthv,p];
            If[Take[v,-overlaptemp]==Take[u,{p-overlaptemp+1,p}],
                match=1;overlap2=overlaptemp,
                i++
            ]
        ]
    ];

    If[overlap1>overlap2,
        distance=(lengthu+lengthv-2*overlap1);
        rdistance=distance/(lengthu+lengthv),
        distance=(lengthu+lengthv-2*overlap2);
        rdistance=distance/(lengthu+lengthv)
    ];

    Print["N+M = ",lengthu+lengthv," ; ","Lmax = ",
        Max[overlap1,overlap2]];
    Print["d = ",distance," ; ","dr = ",rdistance," ; ",
        "p = ",1-rdistance];

    ];
```

Examples

In[133]:=
```
HBTdistance[{1, 1, 2, 3, 5, 8, 13, 21, 34, 55}, {1, 1, 2, 3, 5, 8, 13, 21, 34, 55}]
```

N = 10 ; M = 10

N+M = 20 ; L_{max} = 10

d = 0 ; d_r = 0 ; p = 1

In[134]:=
```
HBTdistance[{1, 1, 2, 3, 5, 8, 13, 21, 34, 55}, {1, 1, 2, 3, 5, 8, 13, 21, 47, 55}]
```

N = 10 ; M = 10

N+M = 20 ; L_{max} = 0

d = 20 ; d_r = 1 ; p = 0

In[135]:=
```
HBTdistance[{1, 1, 2, 3, 5, 8, 13, 21, 34, 55}, {55, 89, 144, 233, 377, 610}]
```

N = 10 ; M = 6

N+M = 16 ; L_{max} = 1

d = 14 ; $d_r = \frac{7}{8}$; $p = \frac{1}{8}$

In[136]:=
```
HBTdistance[{1, 1, 2, 3, 5, 8, 13, 21, 34, 55}, {3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377}]
```

N = 10 ; M = 11

N+M = 21 ; L_{max} = 7

d = 7 ; $d_r = \frac{1}{3}$; $p = \frac{2}{3}$

In[137]:= **HBTdistance**[{**1, 1, 2, 3, 5, 8, 13, 21, 34, 55**}, {**2, 3, 5, 8, 13, 21, 34**}]

N = 10 ; M = 7

N+M = 17 ; L_{max} = 7

d = 3 ; $d_r = \frac{3}{17}$; $p = \frac{14}{17}$

EUREKA Project

- **Database**
- **Over one million sequence transformations (T1-T11) have been calculated (A000001-A170000)**
- **MySQL database of transformed sequences contains over 77 million rows (each row stores a window of 3 terms of a sequence) - 5 GB file**
- **Search Results**
- **Over 300,000 matches found so far ($d_r \leq 1/2$, $L_{max} \geq 4$)**
- **Preliminary analysis shows:**
  - Most matches are trivial or already mentioned in OEIS (> 99%)
  - Small fraction of false positives (> 0.9%)

Ten Experimental Conjectures

■ EUREKA Database Website

1. 1563: A000129S1T3 = A041011S1T8
2. 2010: A000240S1T7 = A006882S1T8
3. 2020: A000241S1T8 = A028723S1T8
4. 2443: A000295S1T9 = A031878S1T4
5. 4850: A001076S1T3 = A041143S1T8
6. 25802: A014445S1T3 = A001076S1T8
7. 56759: A041041S1T3 = A162671S1T8
8. 103439: A108099S1T7 = A132344S1T8
9. 109026: A120580S1T2 = A024493S1T9 (Hankel Transform) A161937S1T7
10. 129200: A161937S1T7 = A087299S1T8

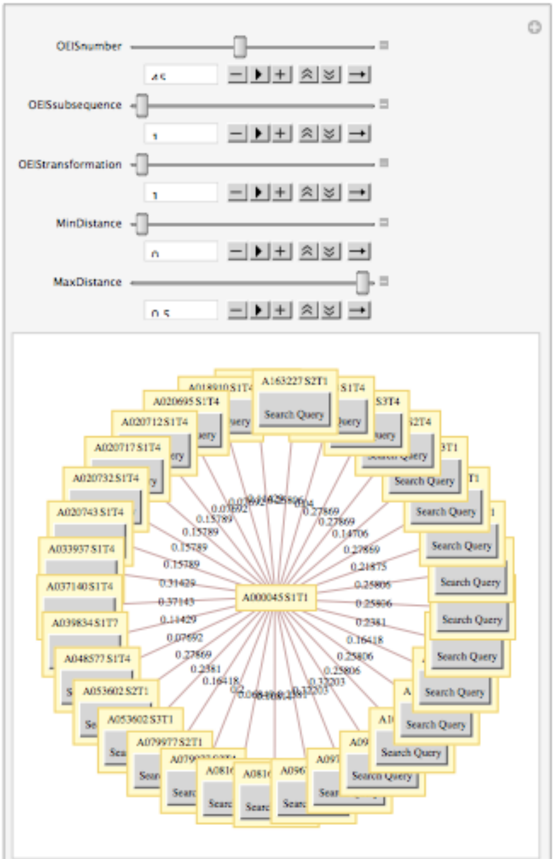
Next Steps

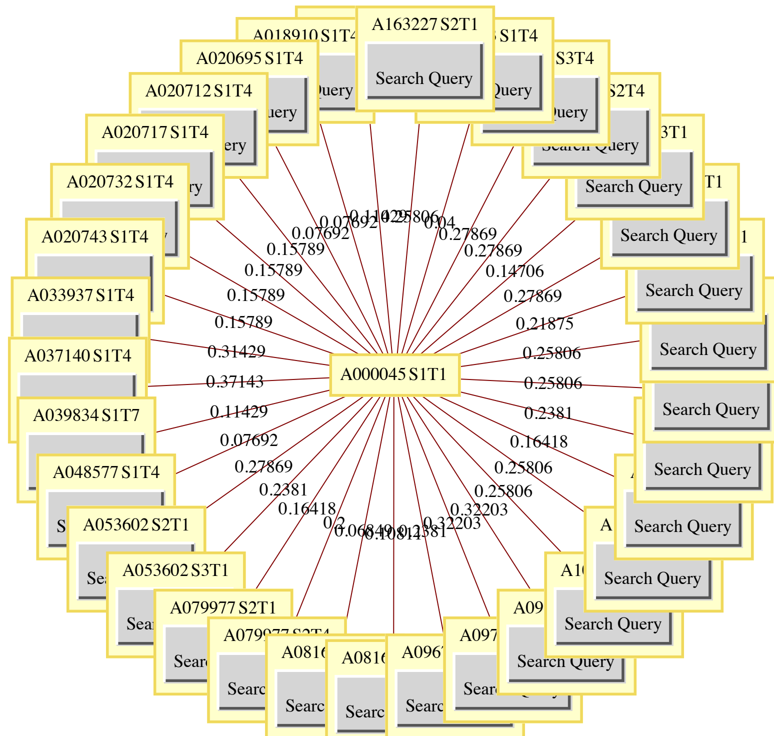
- **Scale up processing power and memory**
- **Perform search on a cluster of computers ✓**
- **Implement parallel/distributed computing (Rowan’s 3-node CC cluster)**
- **Improve sequence matching algorithms**
- **Reduce search-times ✓**
- **Reduce trivial matches and false positives**
- **Expand Scope of Search**
- **Enlarge collection of sequence transformations ✓**
- **Composition of sequence transformations**
- **Extend search to 2-D sequences (e.g. Pascal’s triangle) and rational sequences (e.g. Bernoulli numbers)**

- **Disseminate Work**
- **Create database website ✓**
- **Make database website accessible to the public (collaborate with OEIS)**
- **Graph Network Visualization of Identities**

```
In[15]:= OEISIdentitiesGraphPlot["A000045S1T1"]
```

Out[15]=





- **Publish new interesting (non-trivial) EUREKA's experimental conjectures**
- **Seek Help**
- **Need good programmers (recruit students! ✓)**
- **Need collaborators to analyze and prove EUREKA's experimental conjectures (suitable as student research projects)**

The End

Thank you