OEIS Cross-Reference Network Visualization and PageRank Analysis

Interactive Visualization, Spring 2020

Owen Durkin Owen Xie

May 1, 2020

Abstract

We set out to visualize the On-Line Encyclopedia of Integer Sequences as a network using the cross-references between sequence pages. The visualization extends previous visualization work by computing the PageRank of each sequence as a measure of "importance". As part of this paper, we discuss the results of our PageRank analysis as well as feedback received on the visualization. We hope that this work may serve as a starting point for an interesting way to navigate the OEIS, or that it may be integrated into existing work.

1 Introduction

1.1 The On-Line Encyclopedia of Integer Sequences

The On-Line Encyclopedia of Integer Sequences is a free-to-use online database that contains comments, equations, and research relevant to a wide variety of number sequences. The website has proved to be a helpful resource for both professional researchers and hobbyist mathematicians alike. [Fou20b]

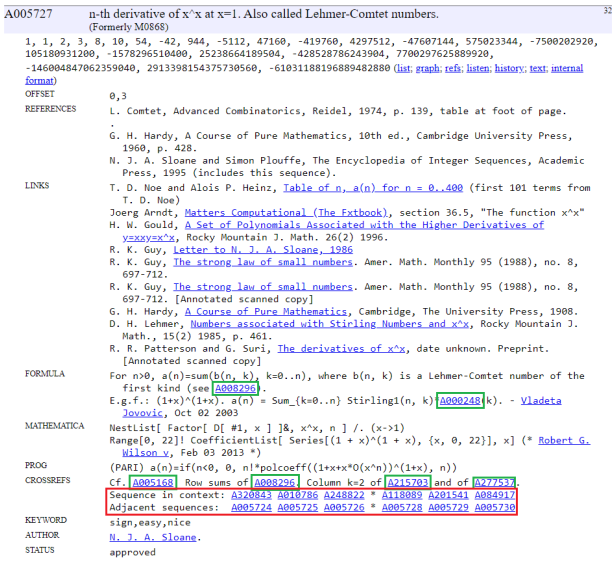
On the website, each page corresponds to a single number sequence with an identifier generated by the letter 'A' followed by 6 digits, coined the "A-number". As shown in Figure 1, a given page of the OEIS may contain references to other sequences in comments and links left by OEIS users. Since these references often indicate some dependency relationship between two sequences, the references generate a network of sequences indicating how sequences contribute to the meaning of other sequences. Accordingly, in this network, each node represents a sequence and each directed edge represents a reference from one sequence to another.

1.2 Visualization and Research Goals

The goal of this project is to visualize the network generated by the cross-references of the OEIS, and hopefully provide an interesting way to navigate the set of integer sequences. We also use the PageRank algorithm to determine which integer sequences are the most

"important", and use that data to inform our visualization design. We initially hypothesized the prime numbers would end up being the highest ranking number sequence. It seems like the mathematicians most obsessed with integer sequences are number theorists, and a lot of number theory is centered around prime numbers and divisibility.

2 Related Work


A005727 n-th derivative of x^x at x=1. Also called Lehmer-Comtet numbers. (Formerly M0868)

1, 1, 2, 3, 8, 10, 54, -42, 944, -5112, 47160, -419760, 4297512, -47607144, 575023344, -7500202920, 105180931200, -1578296510400, 25238664189504, -428528786243904, 7700297625889920, -146004847062359040, 2913398154375730560, -61031188196889482880 (list; graph; refs; listen; history; text; internal format)

OFFSET 0,3

REFERENCES L. Comtet, Advanced Combinatorics, Reidel, 1974, p. 139, table at foot of page.
G. H. Hardy, A Course of Pure Mathematics, 10th ed., Cambridge University Press, 1960, p. 428.
N. J. A. Sloane and Simon Plouffe, The Encyclopedia of Integer Sequences, Academic Press, 1995 (includes this sequence).

LINKS T. D. Noe and Alois P. Heinz, Table of n, a(n) for n = 0..400 (first 101 terms from T. D. Noe)
Joerg Arndt, Matters Computational (The Fxtbook), section 36.5, "The function x^x"
H. W. Gould, A Set of Polynomials Associated with the Higher Derivatives of y=xxy=x^x, Rocky Mountain J. Math. 26(2) 1996.
R. K. Guy, Letter to N. J. A. Sloane, 1986
R. K. Guy, The strong law of small numbers. Amer. Math. Monthly 95 (1988), no. 8, 697-712.
R. K. Guy, The strong law of small numbers. Amer. Math. Monthly 95 (1988), no. 8, 697-712. [Annotated scanned copy]
G. H. Hardy, A Course of Pure Mathematics, Cambridge, The University Press, 1908.
D. H. Lehmer, Numbers associated with Stirling Numbers and x^x, Rocky Mountain J. Math., 15(2) 1985, p. 461.
R. R. Patterson and G. Suri, The derivatives of x^x, date unknown. Preprint. [Annotated scanned copy]

FORMULA For n>0, a(n)=sum(b(n, k), k=0..n), where b(n, k) is a Lehmer-Comtet number of the first kind (see A008296).
E.g.f.: (1+x)^(1+x). a(n) = Sum_{k=0..n} Stirling1(n, k) A000248(k). - Vladeta Jovovic, Oct 02 2003

MATHEMATICA NestList[Factor[D[#1, x]]&, x^x, n] /. (x->1)
Range[0, 22]! CoefficientList[Series[(1 + x)^(1 + x), {x, 0, 22}], x] (* Robert G. Wilson v, Feb 03 2013 *)

PROG (PARI) a(n)=if(n<0, 0, n!*polcoeff((1+x+x*O(x^n))^(1+x), n))

CROSSREFS Cf. A005168, Row sums of A008296, Column k=2 of A215703 and of A277537.
Sequence in context: A320843 A010786 A248822 * A118089 A201541 A084917
Adjacent sequences: A005724 A005725 A005726 * A005728 A005729 A005730

KEYWORD sign,easy,nice

AUTHOR N. J. A. Sloane.

STATUS approved

Figure 1: An example of a page from the OEIS. Cross-references that are accounted for are boxed in green whereas the red-boxed references are ignored. Note that references do not only appear in the "CROSSREFS" section.

Visualization of the OEIS to facilitate discovery and exploration is something that has been attempted before [Abe+16; Str+19]. Abello et al.[Abe+16] focused on defining different topologies of the network for exploration. This project is the most similar to ours in terms of the overall scope, but does not consider the ranking of different sequences.

The question of importance in the OEIS has been considered as well, but on an different scale. In a paper titled "Sloane's Gap: Do Mathematical and Social Factors Explain the Distribution of Numbers in the OEIS?" [GDZ11], Gauvrit, Delahaye, and Zenil investigate a phenomenon, Sloane's Gap, which is a gap in the number of occurrences between "interesting" and "uninteresting" numbers recorded in the OEIS. They proposed that this pattern was a result of both mathematical and human factors. In our case, instead of comparing numbers, we are comparing sequences. We also take advantage of the network topology to generate an importance ranking instead of number of occurances. All the same, this paper inspired us with the general idea that the OEIS itself can be analyzed to draw conclusions about what mathematicians value in mathematics.

3 Design and Implementation

3.1 Initial Design

As seen in the mockup in Figure 2, we planned to have 3 components in our visualization: the network view, the ranking view, and the inspected element view. For the network view, the color of the node would indicate the PageRank of the corresponding sequence. The user can select a node from the network view to bring up the inspected element view, which displays the name and sequence pertaining to a certain sequence.

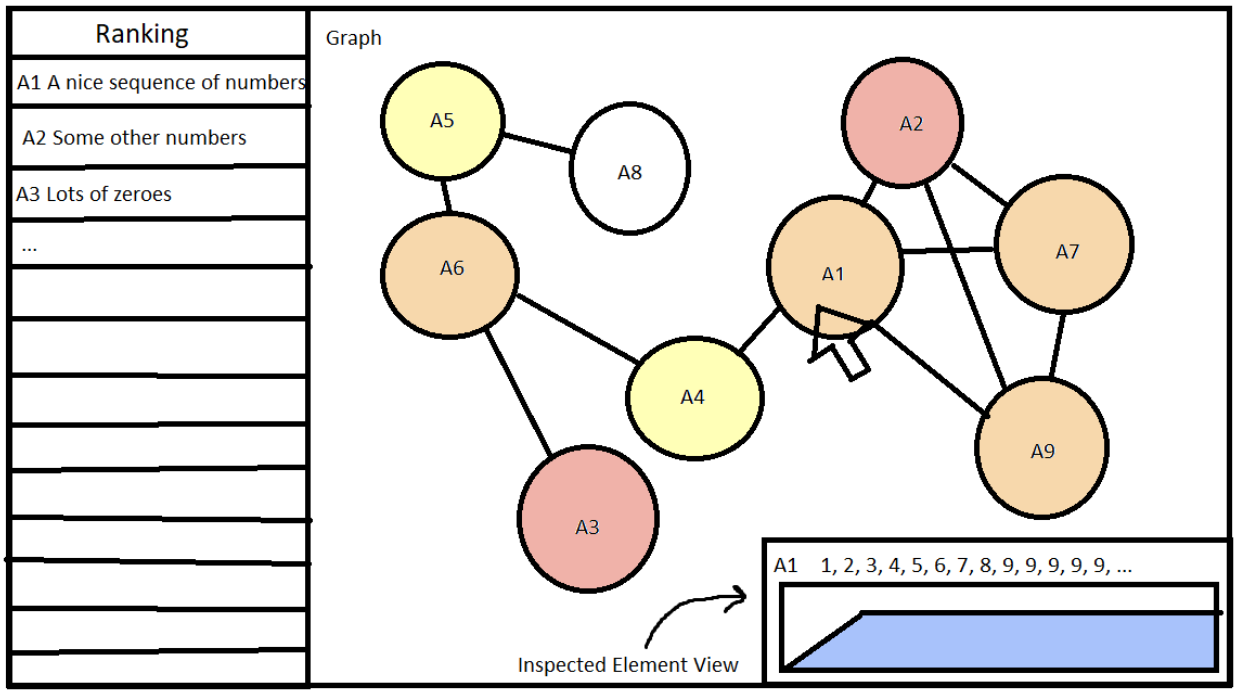


Figure 2: A mockup of the visualization page. The mockup contains 3 components: the network view, the ranking view, and the inspected element view. The saturation of the color of each node indicates its PageRank.

In addition, from peer feedback, we initially did not plan on visualizing the entirety of the OEIS. We anticipated that querying the entire database was going to consume many resources and that visualization with interactive frame rates was near impossible with a large number of nodes in a web application.

While our visualization manages to achieve our goals, we expanded the scope of our project in some regards and curtailed it in others. For example, we initially planned to link the ranking view and network view together, where the user could select a sequence from either the ranking view or the network view, which would highlight the corresponding element of the other view, to give context. In addition, we wanted to add neighborhood highlighting, where the selected node and its neighboring edges would be highlighted, and all other network elements would be dimmed. However, the code we were using for rendering did not include an alpha parameter for rendering nodes and edges, and we did not manage to modify the existing rendering code in time to allow for this alpha-blending. On the other hand, we were eventually able to visualize the entirety of the OEIS with the help of a WebGL graph rendering library, as discussed in later sections.

3.2 Collecting the data

With permission from the president of the OEIS Foundation, Dr. Neil Sloane, we queried each sequence from the OEIS from a text file of sequence names, made available on the website at `https://oeis.org/names.gz`. The python script used to query the data ran in a single thread to avoid making too many requests too quickly. Using the JSON format

of each page accessible via the `fmt=json` parameter, we parsed the entire response with a regular expression (`A[0-9]6`). Each reference that was found in the page was then placed into a NetworkX [HSS08] digraph structure as an edge, where the page's relevant sequence was the source node, while the reference was the target node. Self-loops were removed.

In addition to the references scattered throughout the page, every page contains references to sequences with "close" A-numbers under the `Adjacent sequence` label. The page also contains `Sequence in context` references, which are "close" sequences when the sequences are sorted by their integer contents lexicographically. Based off of initial results containing these sequences, we chose to ignore these references, since they do not necessarily imply a dependency relationship. This way, all of the references in the data set will be those purposefully left by users of the OEIS.

Overall, it took less than 24 hours to collect the data. The final graph was written to two files, an edge list and node list. After all the data had been collected, the data set contained 333,828 nodes and 913,254 edges in total. Of all the nodes, 291,454 nodes had any edges. Singleton nodes are not drawn in our visualization.

3.3 Calculating PageRank

The PageRank algorithm [Win99] was outlined by Page et al. as a method of ranking different web pages. Since the algorithm computes the general idea of page importance, we decided to use the algorithm to rank the integer sequences. Specifically, we used the NetworkX implementation of PageRank, which implements the power iteration method of PageRank.

There are several ways to think about PageRank. In the random surfer model, consider a traveler randomly stumbling through a network. If they reach a dead end, they are moved to a random node in the network. The PageRank of a node is the probability that the traveler is at that node after a very long time. The power iteration model is similar in that it computes those probabilities simultaneously, represented as some vector π. π is then continuously matrix multiplied by a probability transition matrix, which represents the probability that a surfer moves from each node to some other node. This iteration stops when the euclidean distance between π_j and π_{j-1} is smaller than some tolerance, ε.

With NetworkX's PageRank function, we imposed a max iteration count of 10,000 iterations and a tolerance of `1e-12`, which leads to $\varepsilon = 291,454 * 10^{-12}$.

3.4 Building the visualization

The size of the data impacted our choice of tools in visualizing such a large network. From previous experience working with the SVG renderer and d3, we knew that visualizing even around 4,000 nodes would cause a significant slowdown in frame rates. As a result, we decided to utilize the VivaGraphJS graph drawing library [Kas20b], written by Kashcha, which provides a WebGL renderer. Using this renderer, we were able to scale the the visualization up from 50,000 nodes (as seen in Figure 3) to all 300,000 nodes (Figure 4). Using the examples provided, we were also able to change the renderer from using blue squares as a default to colored circles.

One of the many issues with visualizing the large network was the placement of nodes. Force directed layouts are common practice for generating drawings of networks of unknown topology. To avoid computing positions in the browser, we decided to use an offline layout simulator [Kas20a], also written by Kashcha. The offline layout simulates springs between between all pairs of nodes using an force-directed algorithm outlined by Kamada and Kawai [KK89]. Since the algorithm is iterative, we limited the max number of iterations to 1,000 iterations. Overall, this took about 2 hours to run and the results of the layout simulation were placed into a binary file where each triple of 32-bit integers represents the position of the node in that index for the x, y, z dimensions, respectively.

To enable easier sharing of the graph data structure, we used React as a framework alongside Redux, which enables sharing of application state in a more functional way. In addition, to add some basic styling, we used `React-Bootstrap` as our UI framework.

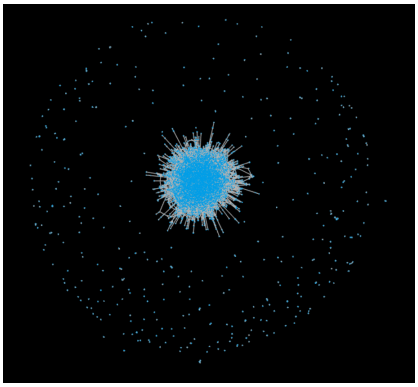


Figure 3: An early prototype of the visualization using VivaGraphJS with 50,000 nodes.

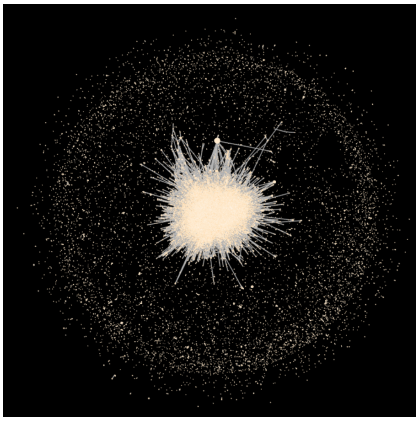


Figure 4: A zoomed out view of the whole network. The nodes with few connections live in an "asteroid belt" around the large cluster in the center. The A250120 cluster is visibly sticking out from the top of the central cluster.

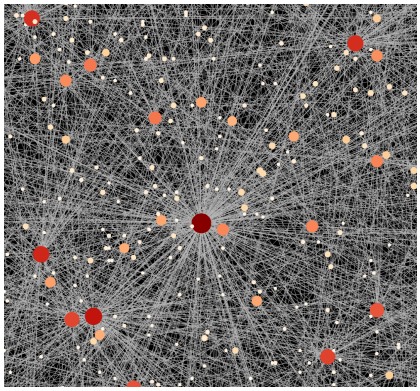


Figure 5: A zoomed in view of nodes. Each node is given a color based on a continuous sequential color scale.

3.5 Visualization Features

3.5.1 Node size and color

To visualize the ranking, the visualization uses the PageRank is to determine the radii and color of nodes in order to emphasize higher ranking sequences. We calculate a score for each node as follows:

$$score(n) = \ln\left(rank_n \times \frac{1}{\min_i(rank_i)}\right)$$

Note that the argument to ln is scaled in such a way that each score is guaranteed to be greater than or equal to 0. To calculate the radius of a node, we interpolate the score over the range $[1,\ \max_n(score(n))]$. Similarly, the color of a node is determined by the score

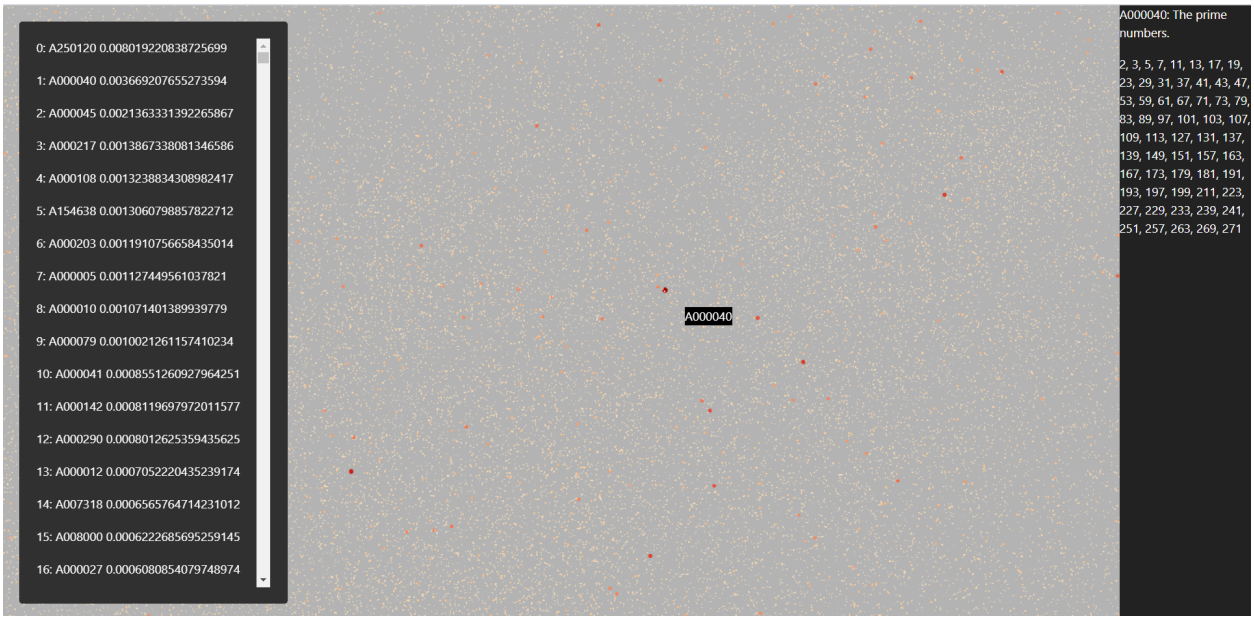


Figure 6: A screenshot of a mouse-over of the prime numbers in the application. In the inspected element view on the right, some information is displayed about the prime numbers. The prime numbers are near the center of the large cluster, so there are tons of nearby nodes. There are so many edges that the entire background is grayed out, which is one motivation for implementing neighborhood highlighting and/or visualizing the network in 3D.

interpolated over a continuous sequential color scale. We used `d3-scale-chromatic` for interpolating colors. We chose to use a log scale because there is a large disparity between the minimum and maximum PageRank, and the majority of nodes have a low PageRank.

3.5.2 Tooltip

On mouse-over of a node, a tooltip appears with the A-number for the sequence represented by that node.

3.5.3 Inspected Element View

On clicking a node, a request is made to the OEIS to display some basic information about that sequence (its name and first few entries).

3.5.4 Ranking View

On the left, a scrollable list of the sequences is sorted by their PageRanks. Next to each sequence, its PageRank is listed. Since loading all of the rankings would slow down the webpage, we made use of the `react-window` library [Vau20] to lazy load of the ranking of the sequences in the ranking view.

4 Analysis of Results

4.1 Top Ranking Sequences

Here we list the top 10 ranking sequences and their A-numbers

1. A250120 Coordination sequence for planar net 3.3.3.3.6 (also called the fsz net).
2. A000040 Prime numbers
3. A000045 Fibonacci numbers
4. A000217 Triangular numbers
5. A000108 Catalan numbers
6. A154638 The number of distinct reduced words of length n in the Coxeter group of "Apollonian reflections" in three dimensions
7. A000203 $\sigma(n)$, the sum of the divisors of n
8. A000005 $d(n)$, the number of divisors of n
9. A000010 $\varphi(n)$, the Euler totient function
10. A000079 2^n, Powers of 2

There are many popular, simple, and important sequences here, which appear again and again in combinatorics and number theory, but the top-ranking sequence certainly stood out to us. What gives?

4.2 The A250120 Cluster

Figure 7 depicts A250120 as it appears in our layout [Fou20a]. The large red node, A250120, is surrounded by some hundreds of smaller nodes, each with a link to it. On further investigation, we found that each of the pages of these sequences contains a link titled "`k-uniform tilings (k <= 6) and their A-numbers`" to the page "https://oeis.org/A250120/a250120.html". So, our regular expression search picked up "A250120" on every page with this link.

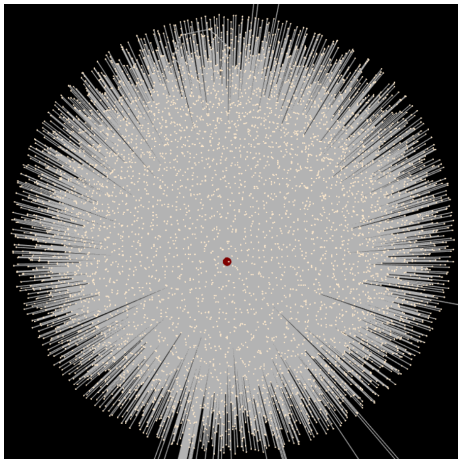


Figure 7: The A250120 cluster. The A250120 sequence is the large red node in the center, and all of the smaller, lighter nodes are the coordination sequences that have a reference to it.

This page classifies uniform tilings of the 2-D plane. A tiling fills the plane with shapes in a repeating pattern that can extend indefinitely, as shown in Figure 8. The page also lists the relevant coordination sequences for each tiling. A coordination sequence is defined by the number of vertices, $a(n)$, of distance

n from a given vertex. The first tiling listed is of the hexagon-net (i.e. honeycomb pattern), whose coordination sequence is A008486. It is safe to assume that all of the sequences listed on the classification page have this link, and are the sequence making up this cluster. The majority of nodes in the cluster are of the form Gal.u.t.v, which "denotes the coordination sequence for a vertex of type v in tiling number t in the Galebach list of u-uniform tilings" (quoted from the OEIS). It seems that the enumeration of k-uniform tilings for all $k \leq 6$ is thanks to work done by OEIS user Brian Galebach.[Gal20]

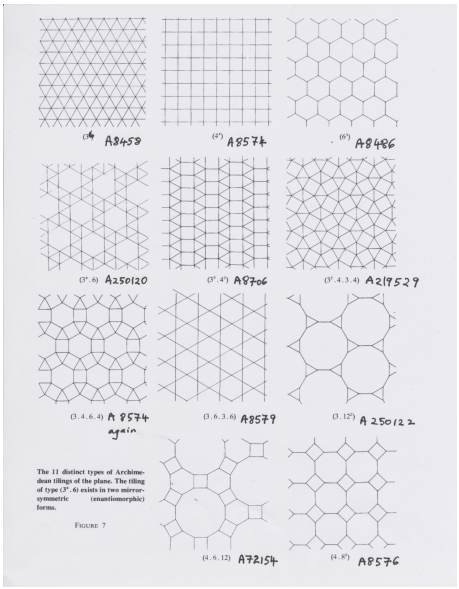


Figure 8: An image of tilings annotated by their relevant coordination sequences posted by N.J.A. Sloane on the page for A008576[Fou20c].

The existence of this cluster was a surprising result to us, and it is an example of how the structure of the OEIS is influenced by human factors. If the list of k-uniform tilings was put under a url not containing an A-number, then it is likely that this cluster would not exist at all, and all of the G.u.t.v coordination sequences would be scattered about the "asteroid belt". Given the nature of how A250120 came to be the highest ranking sequence, it is safe to say that the sequence itself is probably not the most important integer sequence in mathematics. We still believe that title should be reserved for the primes. All the same, this result introduced us to an unfamiliar but beautiful area of mathematics.

4.3 Peer Feedback

Many of our peers described our visualization as "cool".

Most of the criticism of the visualization was concerned with usability. In it's current state, it is not immediately clear for a novice user how the application is intended to be used. Some features (discussed in "Future Work") would make the application significantly more convenient to use. Peers also suggested adding guides for novice users, as well as a reverse lookup capability. Some also suggested future work to improve the frame-rate of the application.

5 Conclusions

In this paper, we have described how we used available technologies to build a novel visualization of the On-Line Encyclopedia of Integer Sequences. In determining the "importance" of each sequence, our top ranking sequence turned out to be an example of how the references on the OEIS are simultaneously necessitated by mathematics and the product of human factors.

5.1 Future Work

There are many hypothetical features we could add to our application to make it more usable. For example, the initial idea of neighborhood highlighting- on selecting a node, the neighborhood is highlighted, and the non-selected nodes are dimmed. It would be helpful to allow users to look up where a sequence is in the layout according to its A-number. The inspected element view could list in-edges and out-edges, with links for navigating to its neighbors. The camera could pan to selected nodes/neighborhoods. Another data set that could be calculated and incorporated into the inspected element view would be the "personalized" PageRank [Hav03] of each node, which lists the importance of nodes relative to a particular node.

In addition, we could experiment with different layouts to try to remedy the issue with edges in the center of the network. Upgrading our visualization to 3D would give the nearly million of edges an extra dimension of breathing room. Edge bundling techniques could also reduce the mess of edges in the middle.

The ranking view could be touched up to be more user friendly. Instead of communicating the raw PageRank, this information could be encoded in a horizontal bar chart. Additionally, the name of the sequence could be included next to the A-number to help identify the sequence.

Finally, Dr. Neil Sloane has expressed interest in this project as a potential addition to the OEIS, to allow users to explore relationships between sequences. Since the project by Abello [Abe+16] had also been in contact with him, we would be working on integrating our PageRank results into the project.

5.2 Division of Labour

Owen Durkin worked primarily on data collection and the inspected element feature, whereas Owen Xie handled the graph layout, graph rendering, and the overall UI setup.

Special Thanks

We would like to thank Dr. Neil Sloane for permitting us to collect the data from the OEIS, and Professor Cutler for convincing him that we are not spammers.

References

[Abe+16] James Abello et al. *Mapping the OEIS*. 2016. URL: https://reu.dimacs.rutgers.edu/2016/proposed.html.

[Fou20a] OEIS Foundation. *Sequence A250120*. 2020. URL: http://oeis.org/A250120.

[Fou20b] OEIS Foundation. *The On-Line Encyclopedia of Integer Sequences*. 2020. URL: http://oeis.org.

[Fou20c] OEIS Foundation. *The uniform planar nets and their A-numbers [Annotated scanned figure from Gruenbaum and Shephard (1977)]*. 2020. URL: http://oeis.org/A008576/a008576.png.

[Gal20] Brian Galebach. *k-uniform tilings (k <= 6) and their A-numbers*. 2020. URL: http://oeis.org/A250120/a250120.html.

[GDZ11] Nicolas Gauvrit, Jean-Paul Delahaye, and Hector Zenil. *Sloane's Gap. Mathematical and Social Factors Explain the Distribution of Numbers in the OEIS*. 2011. arXiv: 1101.4470.

[Hav03] T. H. Haveliwala. "Topic-sensitive PageRank: a context-sensitive ranking algorithm for Web search". In: *IEEE Transactions on Knowledge and Data Engineering* 15.4 (2003), pp. 784–796.

[HSS08] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11–15.

[Kas20a] Andrei Kashcha. *ngraph.offline.layout*. 2020. URL: https://github.com/anvaka/ngraph.offline.layout.

[Kas20b] Andrei Kashcha. *VivaGraphJS*. 2020. URL: https://github.com/anvaka/VivaGraphJS/.

[KK89] Tomihisa Kamada and Satoru Kawai. "An algorithm for drawing general undirected graphs". In: *Information Processing Letters* (1989).

[Str+19] Katherine Strange et al. *Visualizing Integer Sequences, Spring 2019*. Jan. 2019. URL: https://www.colorado.edu/math/visualizing-integer-sequences-spring-2019.

[Vau20] Brian Vaughn. *react-window*. 2020. URL: https://github.com/bvaughn/react-window.

[Win99] Lawrence Page and Sergey Brin and Rajeev Motwani and Terry Winograd. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999-66. Previous number = SIDL-WP-1999-0120. Stanford InfoLab, Nov. 1999. URL: http://ilpubs.stanford.edu:8090/422/.