A COMPLEX NETWORKS APPROACH TO ANALYSING THE ERDŐS-STRAUS CONJECTURE AND RELATED PROBLEMS

VIKHYATH MONDRETI

STUDENT, PIONEER ACADEMICS RESEARCH PROGRAM, 104 S 20TH ST., PHILADELPHIA, PA 19103, USA

Abstract. For any positive integer $n \geq 2$, the *Erdős-Straus Conjecture* claims that the Diophantine equation $\frac{4}{n} = \frac{1}{x} + \frac{1}{y} + \frac{1}{z}$ has a solution where x, y, z are also positive integers. In this paper, a directed network based on this equation is generated, with properties such as its average clustering coefficient, average path length, degree distributions, and largest strongly connected component analysed to reveal some underlying trends about the nature of the conjecture. Potential connections between different numbers, that result from satisfying a source-solution relationship for this equation, are described using the appropriate number-theoretic interpretations wherever possible, while conjectures backed by these trends are made in other instances. Additionally, a directed configuration model is used to show that the origin of several results is the degree sequence of the network. Metrics relating to the prime number nodes, specifically their in and out degrees, are also explored to yield some intriguing observations. On the whole, the aim is to highlight the viability of complex networks as a computational tool to study this general class of problems pertaining to fixed-length unit fraction splits.

Keywords: Degree Sequence, Directed Networks, Egyptian Fractions, Erdős-Straus Conjecture

1. Introduction

The utilisation of networks as a tool for computational analysis has undoubtedly become commonplace in a range of fields both in the physical and mathematical sciences as well as the social sciences. The inherent ability of networks to provide a large-scale overview of highly intricate systems simply based on fairly objective 'relationships' between their elements is effectively what makes them so versatile and more importantly, applicable to the real world. This paper, through the various observations presented, will attempt to not only provide a different perspective of a long-standing number theory conjecture but in the process display the usefulness of a complex network representation as a heuristic approach.

The conjecture, formulated by Paul Erdős and Ernst G. Straus in around 1950 [6], states that the following Diophantine equation has a positive integer solution triplet (x, y, z) for all integers $n \geq 2$:

$$\frac{4}{n} = \frac{1}{x} + \frac{1}{y} + \frac{1}{z}$$

It is important to note that this paper will assume that the solutions are distinct, thereby forming a three-term *Egyptian fraction* representation of $\frac{4}{n}$, and are in increasing order. This helps in making the computations much faster as fewer solutions need to be found. However, the nature of the problem is not affected as: if there exists any solution triplet, then there must also exist a solution triplet with distinct integers for $n \geq 3$ (n = 2 is an exception with a trivial known solution) [5]. Considering that the conjecture solely requires that at least one solution exists for the equation, this assumption has previously been made by several researchers. Additionally, it is clear that the conjecture only needs to be solved

for the primes as all positive integers (except the unit 1) are multiples of some prime, and therefore must have solution triplets in which x, y, and z are also the same multiple of their prime divisor's triplets. This is exactly why the prime number nodes in this paper's network representation are analysed independently as well.

For some more context, the *Erdős-Straus Conjecture* is to date unproven. Analytical proofs have achieved significant partial results with a focus on solving for n in different residue classes. The strongest results using this method were obtained by Mordell [12] who proved the conjecture for all n except for primes congruent to $1^2, 11^2, 13^2, 17^2, 19^2,$ or $23^2 \bmod 840$ which could thereby hold counterexamples. Apart from being used in known sieve theoretic methods, these soluble classes have also supported efficient computational verification, with Swett [19] verifying it for $n \leq 10^{14}$. A similar conjecture for $\frac{5}{n}$ was made by Sierpiński [17] with the more generalised version of this problem class coming from Schinzel [14].

2. Model

In this section, a simple directed network model, briefly termed the Erdős-Straus Conjecture (ESC) graph, used to represent the Diophantine equation is introduced. The function $S(n)$ outputs an unordered set of integers that appear as part of any of the solutions to the equation. $D(n) = \{k \in \mathbb{N} : k \leq n\}$ and is the set containing all the denominators that have been solved for. The graph itself is denoted by $G_{ESC}(n)$ with its vertex and edge sets being defined as follows.

Vertex Set:

$$V(n) = D(n) \bigcup_{i \in D(n)} S(i)$$

Edge Set:

$$E(n) = \{(s,t) : \forall t \in S(s) \ for \ each \ s \in D(n)\}$$

Essentially, $G_{ESC}(n)$ represents a simple directed network where all the denominators that have been tested produce directed edges to all their unique solution integers, even if it includes themselves (self-loops), in each of the respective triplets. This implies that the tested denominator always has an out-degree greater than or equal to 3 unless a counterexample is found. $S(n)$, which is fundamental to find the target nodes, requires all the Egyptian fraction splits to be found for any denominator. There are multiple construction algorithms that can be used to do this including the relatively popular Fibonacci-Sylvester Greedy algorithm and Golomb's method [9]. However, in this paper an open-source Python implementation [8] of Dr. Ian Stewart's Short Sequence Method for recursive brute-force [18] is used as it proved sufficient for all computations done here.

There are a number of ways this equation could have been modelled as a network, with undirected graphs and directed multigraphs both options, as in the case of the latter, the number of times a solution integer is referenced by an n (source denominator) can also be counted via the parallel edges. However, to simply observe and track the unique n-solution relationships between different integers, this representation was deemed adequate, for reasons of more efficient computation and possibly, reduced mathematical 'noise' that will make analysis easier without affecting any fundamental conditions of the conjecture itself.

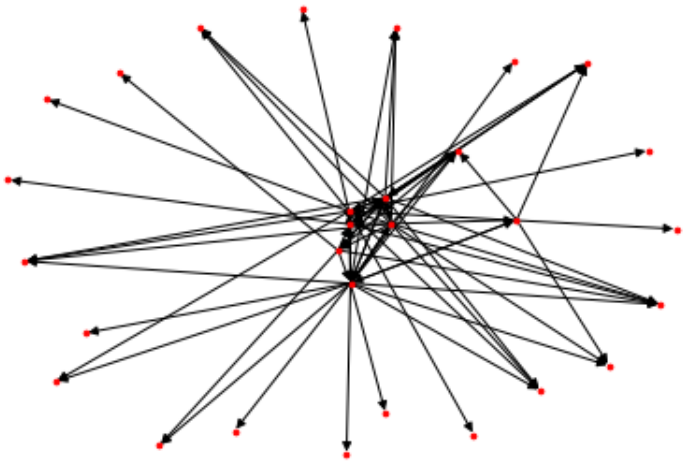

FIGURE 1. $G_{ESC}(10)$ with 29 nodes and 79 edges

3. Degree Distribution

The degree distribution of the network is an interesting metric to look at as it offers a holistic glimpse into its potentially scale-free nature. In this case, the in-degree sequence would be a more logical choice as it would reveal the tendency of certain numbers to be solutions to the equation more often than others - thereby, possibly yielding the extent of preferential attachment embedded into the mathematical depth present.

It is also important to note at this point that the Largest Strongly Connected Component[1] (SCC), almost always contains all source-denominator nodes (solved n-values). This is a highly useful observation as it implies that as the number of n-values tested tends to infinity, which would be the required case for proving the correctness of the conjecture, all positive integers become a part of the largest SCC. Therefore, analysing this as a sub-graph might help gain insight into the properties of the network that might be present at $G_{ESC}(\infty)$ - and will therefore be a recurring tool in this paper.

Scale-free networks, whose degree distribution follows the power law [1], usually have

$$P(k) \sim k^{-\gamma} \ (where\ 2 < \gamma < 3)$$

From *Figure 2*, it is clear that the overall network has some level of preferential attachment with $\gamma \approx 1.80$. However, this apparent preference for some solution integers over others seems to disappear in the largest SCC as $\gamma \approx 0.546$ in *Figure 3*. The reason for this observation is simply the existence of a finite number of integers with a high number of factors versus an infinite number when n-values also tend to infinity. This is because the conjecture can effectively be restated as

$$\frac{4}{mp} = \frac{1}{mx} + \frac{1}{my} + \frac{1}{mz} \ (where\ p\ is\ any\ prime)$$

Similar to the reason for why the conjecture needs to be solved only for the primes - the solution sets for composites are simply multiples of the solutions sets of their factors. Therefore, the more number of factors a number has, the more likely it is to show up as a solution (and in fact, even a source i.e. high out-degree) as it would simply be a different multiple of some

[1]The Largest SCC and effects of the number of factors are further explained in the *Other Trends* section.

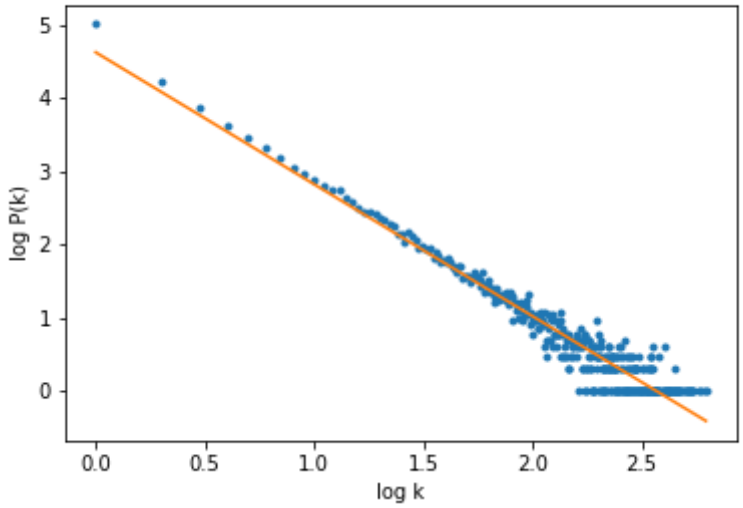

FIGURE 2. In-degree Distribution for $G_{ESC}(1000)$

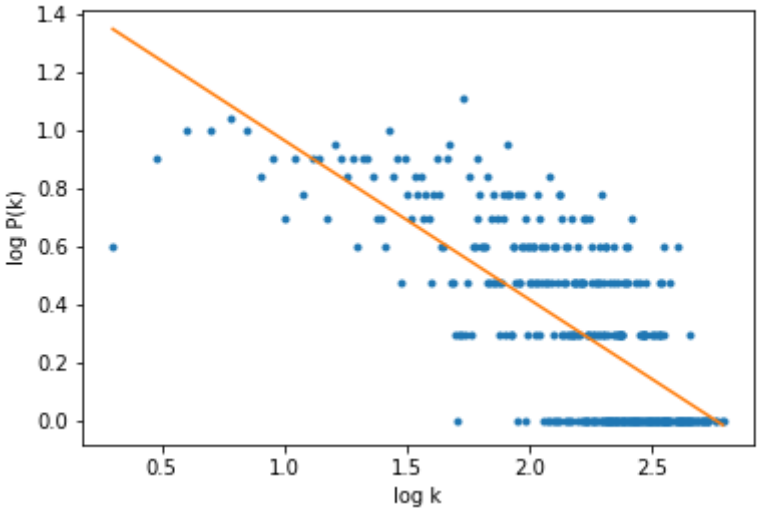

FIGURE 3. In-degree Distribution for the Largest SCC of $G_{ESC}(1000)$

base solution triplet. This justifies the preferential attachment when $G_{ESC}(n)$ is computed for a relatively small n like 1000. However, as this $n \to \infty$ these high-factor numbers are no longer the 'centres' but instead the probability of being a solution also spreads out amongst the now abundant (or ideally, infinite) number of high-factor integers. The in-degree distribution is consequently imperative to understanding many other core metrics of the network.

4. Average Shortest Path Length

The average shortest path length (directed) was computed using the standard formula,

$$a = \sum_{s,t \in V(n)} \frac{d(s,t)}{x(x-1)} \qquad (where\ x\ is\ the\ number\ of\ nodes)$$

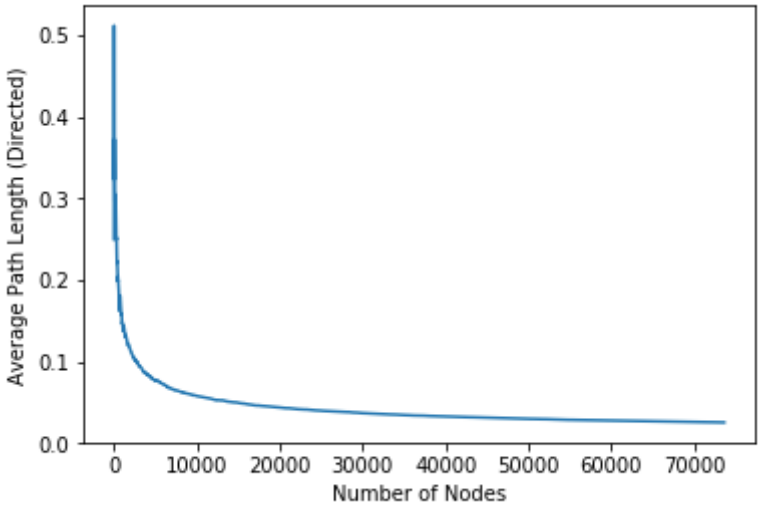

FIGURE 4. Average Path Length (Directed) against the Number of Nodes

The value is < 1 because in the full network because most nodes are only present as they are solutions and have not themselves been solved for as sources, therefore resulting in shortest path lengths of 0 to all other nodes by default. Consequently, *Figure 4* can be considered as a mere artifact of computing the average path length for directed graphs. Although this makes the largest SCC sub-graph far more pertinent, one intriguing non-trivial deduction that explains the exponential decrease in the average path length with the growth of the network is that the number of solutions which are a lot larger than the source are also exponentially more frequent than are solutions less than the source node being solved (as those would have non-zero path lengths). Since this must be reflected within the degree sequence of the network, a random network based on this degree-sequence was generated using a directed configuration model [13]. As expected, the average path length did appear to have its foundations here as the random network yielded an identical average shortest path length to the actual $G_{ESC}(1000)$ network (value not considered in *Figure 4*).

For the largest SCC, however, the value appeared to be approaching 2.0 even though a rigorous proof for stabilisation could not be deduced. In *Figure 5* the largest components calculated are until approximately $G_{ESC}(600)$, enough to judge the general trend, although by $G_{ESC}(1000)$ the value is almost exactly 2.0, and 2.03 by $G_{ESC}(2000)$ indicating slowing growth and hence possible future stabilisation. This might also be of interest as it implies that once again, as n-values tend to infinity almost every number would be able to reach another within about two edges. Although this paper will not look to expand on any analytical mathematical proofs, this seemingly incredible answer probably does have some intrinsic tie with the previously stated fact that almost all solved n-values are part of the largest SCC immediately (with at most minor delays for some primes).

Future study on this relation between the solutions of a specific n and the solutions of these solution integers themselves may help speed up Egyptian fraction splits for this conjecture as the prediction can be based on the fact that the number should only be at most a fixed distance away (here approximated to be 2.0) in this network representation for sufficiently large n-values, possibly for some form of a more computationally cheaper brute-force.

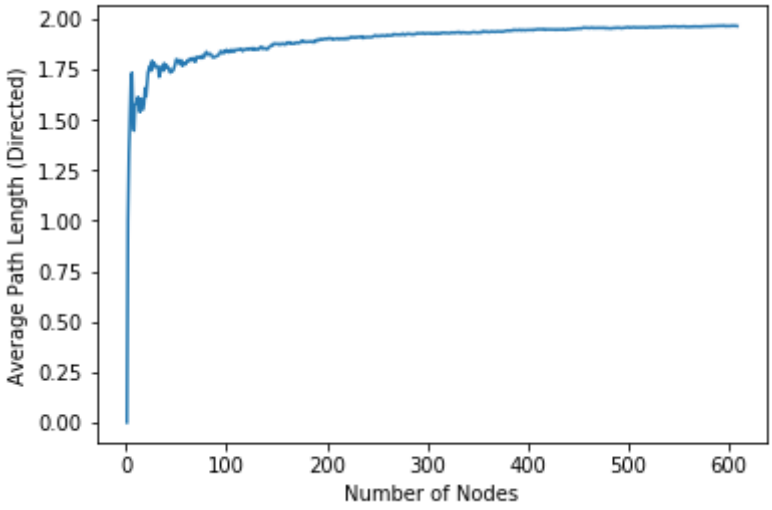

FIGURE 5. Average Path Length (Directed) against the Number of Nodes in Largest SCCs

Furthermore, the largest SCC of the random degree-sequence based network was in fact very similar with five nodes less, even though it had about 27000 edges less and a significantly lower average degree as a result. Most interestingly, it too had an average shortest path length of almost exactly 2.0 again reiterating the point that the degree sequence is the likely reason behind most path-length properties observed.

5. Average Clustering Coefficient

The clustering coefficient (directed) was computed using the formula [2],

$$c_u = \frac{1}{deg^{tot}(u)(deg^{tot}(u) - 1) - 2deg^{\leftrightarrow}(u)} T(u)$$

where $T(u)$ is the number of directed triangles through node u and $deg^{tot}(u)$ is the sum of in-degree and out-degree of u and $deg^{\leftrightarrow}(u)$ is the reciprocal degree of u.

Averaging it out for each node,

$$C = \frac{1}{x} \sum_{v \in G_{ESC}(n)} c_v \qquad (where\ x\ is\ the\ number\ of\ nodes)$$

To explain the clustering present in the network (only the largest SCC would be of relevance here as others would not have out-degrees), the most plausible approach was to observe the numbers forming these directed triangles. Via a triadic census [21] on the network, the 030C type triads (directed triangles) were sampled and manually analysed. Once again, similar to the path-length observations, the triangles appeared to be constructions of a probabilistic model, as when the n and the solution nodes share a prime factor, there exists the greater chance of forming the edge. When these node pairs sharing prime factors form a loop of three, it happens to complete a directed triangle. The key, however, lies in a second trend - no directed triangle with all three nodes as primes was found. Although it is possible for a prime to have a prime solution, it was unclear as to why these loops did not form with all three primes or whether the sample sizes taken were simply not sufficient, in which case, the rarity of these occurrences would be a point of note.

This, however, would suggest that the average clustering, which references the fraction of complete triangles, can be estimated by using the prime count [23]. For $n > 10$, a reasonably accurate estimate was found to be

$$C_{estimate} = \frac{2\pi(n)}{n} \qquad (where\ \pi(n)\ is\ the\ prime\ counting\ function)$$

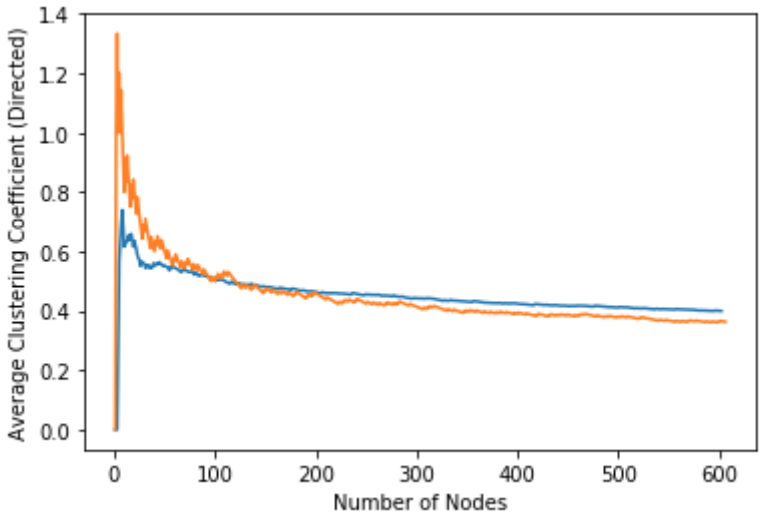

FIGURE 6. C (Blue) compared to the $C_{Estimate}$ (Orange) against the Number of Nodes in Largest SCCs

A central caveat before interpreting this result must be that no concrete mathematical proof or combinatorial argument to explain the triangle formation could be found and that this expression, using the prime number theorem [22], itself becomes

$$\frac{2\pi(n)}{n} \sim \frac{2}{ln(n)}$$

which might make it a co-occurrence as such logarithmic growth can often be seen in a range of models across the natural sciences and mathematics.

6. PRIMES

Considering the importance of primes to this conjecture, and their uniqueness in terms of 'immunity' to any factor-based trends unlike the composites, it was deemed that the metrics of the primes were a major facet for exploration. Specifically, the out-degrees and in-degrees were looked into as they would be representative of the number of solutions source primes have (i.e. when n is prime) and the number of times a prime itself is a solution to the conjecture's equation respectively.

Figure 7 is almost identical with respect to the distribution and structure to the graph obtained by Elsholtz and Tao [4] when counting the number of solutions $f(p)$ for all $p \leq 1000$, indicating that the out-degree is in fact, a fairly reliable metric to monitor this seemingly random number of solutions for the range of primes. The asymptotic bounds proven in the paper were

$$N \log^2 N \ll \sum_{p \leq N} f(p) \ll N \log^2 N \log\log N.$$

A similar scaling for these upper and lower bounds in terms of the out-degree could also possibly be derived although this line of investigation was beyond the scope of this paper.

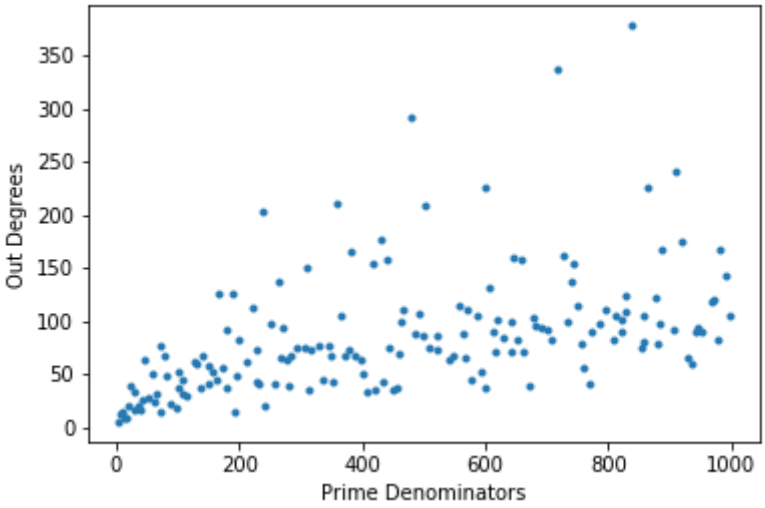

FIGURE 7. Out-degrees against the Source Prime Denominators

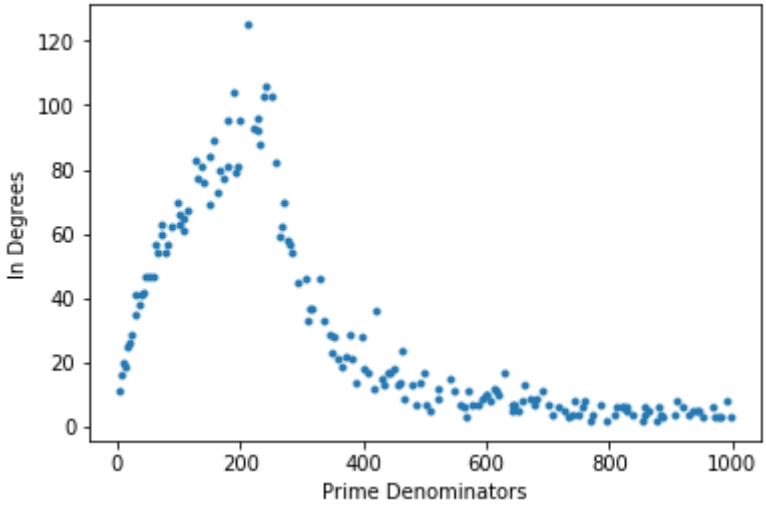

FIGURE 8. In-degrees against the Source Prime Denominators

In-degrees (*Figure 8*), illustrate another notable trend, showing that some primes seem to occur far more often as solutions than others, and that the peak appears to be quite distinct (for $p \leq 1000$, this peak occurs at the prime 211 with an in-degree of 125). The smooth curvature, without too much randomness or outliers from the clearly seen trend, of the graph also seems to point at a deeper underlying cause for this phenomenon. Although this trend might change significantly as more prime denominators are tested, it would still remain an intriguing point of subsequent mathematical study.

7. Other Trends

7.1. **Self-loops.** Considering that self-loops are allowed in this directed network model, when they occur they imply that the Erdős-Straus Dipohantine equation would simplify to

$$\frac{3}{n} = \frac{1}{y} + \frac{1}{z}$$

assuming that x was the solution equal to n, without the loss of generality. It is known that this equation has solutions iff n has a factor congruent to 2 mod 3, with Klee and Wagon [20] attributing this result to Nakayama without a citation. Therefore, self-loops were observed for only these values of n as nodes in the $G_{ESC}(n)$.

7.2. **Largest SCC.** As noted previously, a key observation in this network was the near continuous presence of all solved n-values in the largest SCC barring minor delays in joining the component for some primes. The delays could be partially deduced from the fact that the plot of the number of nodes in Largest SCC against the solved denominators (not shown) was not completely straight (occasional slight curvatures) even though it displayed a strong directly proportional trend on the whole - although the reason as to why the numbers observed with this lag in this range were only primes forms another potential extension of this trend. That this holds from almost the initiation of the node generation process is what is most surprising, and monitoring this for far larger ranges of n might yield other useful results.

7.3. **Growth of Network.** The network, as seen in *Figure 9*, grows on an exponential rate. The reason for this most certainly has a relation with another bound referenced by Elsholtz and Tao [4], and proven by Heath-Brown [11] in a private communication (strongest version of this bound), where the lower bound for the sum of the number of solutions $f(n)$ increases at an exponential rate,

$$\sum_{n \leq N} f(n) \gg N \log^6 N$$

implying similar growth for the number of nodes in the network representation (size) as well, as it too represents an aggregate value like the sum above.

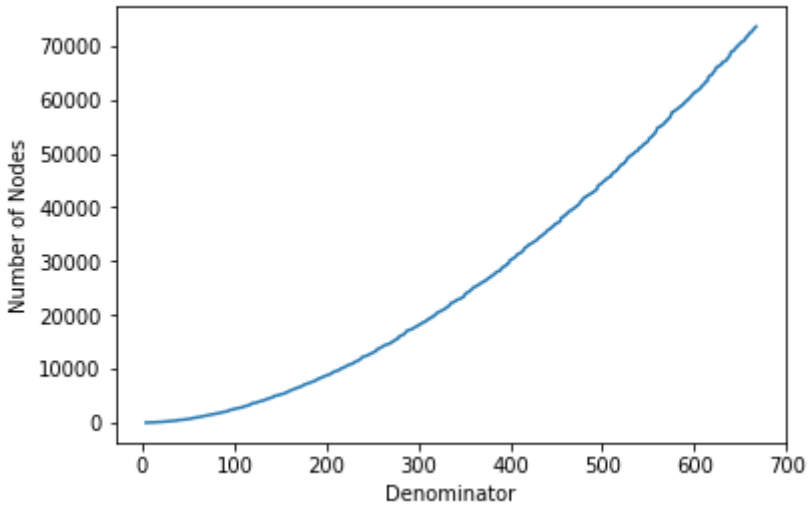

Figure 9. Growth of the Network with the Number of Source Denominators Solved

7.4. **Number of Factors.** The effect the number of factors source nodes have on their number of solutions has been another recurring point in this paper and is critical to putting the other properties of the network into context. As given in the explanation for the degree distribution, it presumably comes down to a high number of factors giving an integer a higher probability of appearing as a source or solution (for out and in degrees respectively). The roughly proportional trend between the number of factors and both the degree sequences is illustrated below for all solved denominators, but not within a separate sub-graph like what was used for the degree distribution computations (thereby making it imperative to note the room for change with a larger range, for the in-degree graph). However, factors certainly cannot be the sole element driving the metrics, with the sequences for primes not affected by this as stated before, and with the same number of factors leading to notably varying out-degrees in general (vertical data-points in the graphs).

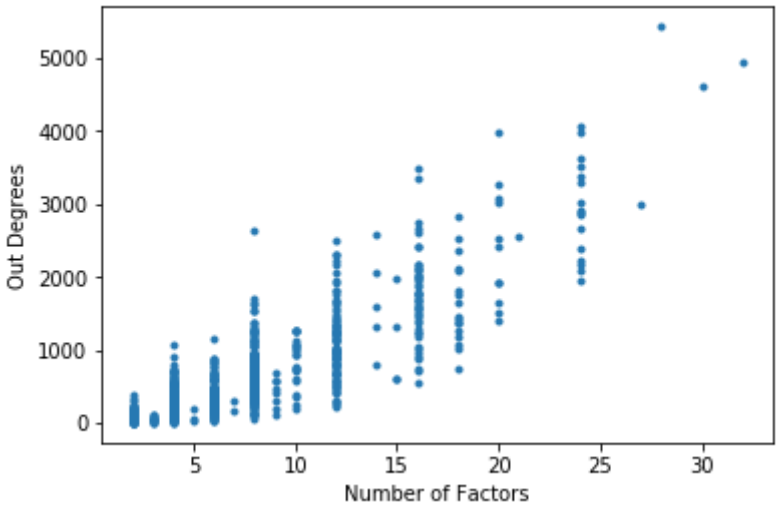

FIGURE 10. Out-Degrees of nodes against their Number of Factors

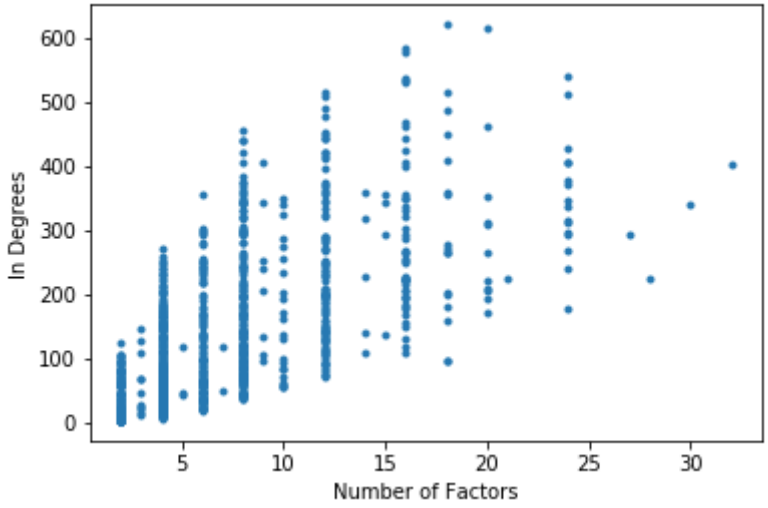

FIGURE 11. In-Degrees of nodes against their Number of Factors

7.5. **Sierpiński's Conjecture.** An identical directed network model was computed for Wacław Sierpiński's variation of this conjecture [17] which states that for all $n \geq 2$, the following Diophantine equation has solutions.

$$\frac{5}{n} = \frac{1}{x} + \frac{1}{y} + \frac{1}{z}$$

Near identical trends to those seen in the Erdős-Straus network were observed in this network too, indicating a certain similarity between the underlying number-theoretic structures of these two equations. The parallel trends included the largest SCC that almost always consisted of all source nodes, degree distributions, average clustering coefficients (in fact, close in raw values as well to $G_{ESC}(n)$), approximately equal average shortest path length when compared to its degree-sequence based random network (also comparable to the $G_{ESC}(n)$ 2.0 value for the largest SCC), the in and out degree prime node trends, the growth rate, and the average effects on the degrees of nodes by the number of factors they have.

8. Conclusion

In the end, the paper provides a range of fascinating insights into this conjecture wrapping back to the main goal of presenting a complex networks representation as a different and useful approach for not only the *Erdős-Straus Conjecture* but also easily extendable to this category of related problems that require fixed-term unit fraction splits. The real value would, however, abide in transcending many of these experimental trends into rigorous mathematical proofs, especially for the several probabilistic arguments that determined most of the analysed metrics.

Network analyses in pure mathematics, and more specifically in number theory, have witnessed rapid growth in recent years in terms of the work done in the study of prime numbers [7], divisibility patterns [15], and a wide array of other integer sequence based networks [3, 16, 24]. As increasingly standardised techniques specialised to this domain are developed to bring greater context, usability, and accuracy to the interpretations offered by these networks, the scope of such research would most definitely be augmented as it can then function in a better known framework more pertinent to the mathematical analysis being conducted.

As for the problem deconstructed here, it was most intriguing to find that the random directed configuration model based only on the degree sequences, and therefore the appearance of these different integers as source nodes or solutions in the Diophantine equation, produced a relatively accurate estimate of the overall network - without any of the mathematical depth evoked by the individual numbers themselves - indicating a greater than expected extent of symmetry and predictability in the number of solutions and the growth of that quantity with n. While many of these trends such as the continuously source-capturing largest SCC, average path length, and the average clustering act as pointers to their own unique properties of the conjecture, perhaps even hinting at its correctness due to the presence of a non-trivial degree of evolving interconnectedness, the answer undeniably has always and will continue to remain amongst the beautiful yet incredibly elusive primes.

References

[1] Barabási, A-L. and Albert, R. "Emergence of Scaling in Random Networks." Science 286, 509-512, 1999.

[2] Clustering in complex directed networks by G. Fagiolo, Physical Review E, 76(2), 026107 (2007).

[3] Corso, G. Families and clustering in a natural numbers network. Phys. Rev. E 69, 036106 (2004).

[4] Elsholtz Christian, and Terence Tao. COUNTING THE NUMBER OF SOLUTIONS TO THE ERDŐS—STRAUS EQUATION ON UNIT FRACTIONS. Journal of the Australian Mathematical Society 94, no. 1 (2013): 50—105. doi:10.1017/S1446788712000468

[5] Eppstein (1995) https://www.ics.uci.edu/~eppstein/numth/egypt/conflict.html. Accessed on 25 August 2019.

[6] Erdős, Paul (1950), "Az 1/x1 + 1/x2 + ... + 1/xn = a/b egyenlet egész számú megoldásairól (On a Diophantine Equation)". Matematiki Lapok. (in Hungarian), 1: 192-210

[7] G. Garcia-Perez, M. A. Serrano and M. Boguna. The complex architecture of primes and natural numbers. arXiv:1402.3612

[8] https://github.com/shaneallgeier/egyptian-fraction-generator. Accessed on 10 August 2019.

[9] Gong, K. "Egyptian Fractions" http://kevingong.com/Math/EgyptianFractions.pdf. Accessed on August 14 2019.

[10] Hardy, G. H.; Ramanujan, S. (1917), "The normal number of prime factors of a number n", Quarterly Journal of Mathematics, 48: 76-92, JFM 46.0262.03

[11] D. R. Heath-Brown, The density of rational points on Cayley's cubic surface, Proceedings of the Session in Analytic Number Theory and Diophantine Equations, 33 pp., Bonner Math. Schriften, 360, Univ. Bonn, Bonn, 2003

[12] Mordell, L. J. Diophantine Equations. London: Academic Press, pp. 287-290, 1969

[13] Newman, M. E. J. and Strogatz, S. H. and Watts, D. J. Random graphs with arbitrary degree distributions and their applications Phys. Rev. E, 64, 026118 (2001)

[14] Schinzel, André (1956), "Sur quelques propriétés des nombres 3/n et 4/n, oú n est un nombre impair", Mathesis (in French)

[15] Shekatkar, S. M., Bhagwat, C. Ambika, G. Divisibility patterns of natural numbers as a complex network. Sci. Rep. 5, 14280 (2015).

[16] Shi, D.-H. Zhou, H.-J. Natural number network and prime number theorem. Complex Systems and Complexity Science 7, 52-54 (2010).

[17] Sierpiński, Waclaw (1956), "Sur les décompositions de nombres rationnels en fractions primaires", Mathesis (in French)

[18] Stewart, I. "The riddle of the vanishing camel." Scientific American, June 1992, pp. 122-124.

[19] Swett, A. "The Erdos-Strauss Conjecture." Rev. 10/28/99.http://math.uindy.edu/swett/esc.htm.

[20] V. Klee and S. Wagon. Old and New Unsolved Problems in Plane Geometry and Number Theory. Math. Assoc. of America, 1991, pp. 175-177 and 206-208.

[21] Vladimir Batagelj and Andrej Mrvar, A subquadratic triad census algorithm for large sparse networks with small maximum degree, University of Ljubljana. http://vlado.fmf.uni-lj.si/pub/networks/doc/triads/triads.pdf

[22] Weisstein, Eric W. "Prime Number Theorem." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/PrimeNumberTheorem.html

[23] Weisstein, Eric W. "Prime Counting Function." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/PrimeCountingFunction.html

[24] Zhou, T., Wang, B.-H., Hui, P. Chan, K. Topological properties of integer networks. Physica A 367, 613-618 (2006).