

# Fitting Random Networks to Road Networks

## Using Greedy Algorithms

Amit Ofek<sup>1</sup> (ae422@scarletmail.rutgers.edu), Meet Patel<sup>1</sup> (mvpatel@princeton.edu), Jürgen Hackl

### ABSTRACT

Research in recreating road networks is often complex and hard to understand without specialized knowledge from the field; many papers can be daunting for newcomers, and the associated models could prove too complex for use whether through their size or application of hard to grasp high-level concepts. Given the lack of easy-to-understand and simplistic algorithms for generating road networks, this paper attempts to use greedy algorithms to create realistic road networks from a given degree of distribution and clustering coefficient while aiming to be simple, understandable, and widely applicable to other similar networks. We create three separate models that each have their own approach: Model 1 picks random nodes from the whole network, Model 2 picks from neighbors of neighbors to create triades, and Model 3 picks from nodes in a manner that aims to maintain the original degree distribution. Incorporating ideas from Markov Chain based approaches, generalized transportation networks, and greedy algorithms, the paper shows that simple models can achieve some of the desired qualities for the model, with Models 1 and 2 producing promising results. While none of the three models proposed entirely fulfilled the objectives that we initially had outlined for a successful model, we believe that the core concepts for each model and for the ideas in the paper could be further developed into one or more successful models that produce acceptable (or better) results while maintaining their simplicity to allow for ease of access.

GitHub code repository: <https://github.com/meetpatel450/cee520-final-project/tree/main>

---

<sup>1</sup> indicates equal contributions

## INTRODUCTION

Networks are everywhere in the real world. Social networks, power grids, and transportation infrastructure dictate how people interact, live, and get from place to place. Studying those networks reveals critical insights that can be used to improve the networks and as a result human lives. Improvements in understanding of power grids, for example, can help reduce the amount of energy production, which in turn, reduces the carbon emissions and aids in improving the state of the environment. Thorough understanding of the road networks provides information on how to best improve roads to increase reliability and resilience while reducing costs; such insights are critical in natural disaster cases where part of the network is shut down and lives depend on the ability to use the roads to transport aid.

A common way to study networks is to create models whose microstates are representative of real-world networks. Two well known examples are the  $G(n,m)$ <sup>1</sup>, and  $G(n,p)$ <sup>2</sup> models. The  $G(n,m)$  model assumes that the number of nodes and the number of links are known. Each microstate of the  $G(n,m)$  model is a different configuration of nodes,  $n$ , and links,  $m$ . This model is very restrictive as it requires a large amount of upfront knowledge about the desired network. The  $G(n,p)$  model requires the number of nodes,  $n$ , and the likelihood of any two nodes having a link in between them,  $p$ . Both  $G(n,m)$  and  $G(n,p)$  are infrequently representative of real world road networks as they exhibit near zero clustering and very low diameters. Since the road networks that are used in this study do not have similar network characteristics, it was decided to explore other models.

Another model that is relevant is the Molloy Reed model<sup>3</sup>. This model creates microstates that have a desired degree distribution. The desired degree distribution is the input to the model and the microstates of the model are different configurations that end up with the degree distribution that was used as an input. Road networks often have specific degree distributions, which would be created in a microstate of the Molloy Reed if the degree distribution is known. However, the Molloy Reed microstates have much lower clustering coefficient than road networks. To address this, our model will need a way to improve the clustering coefficient.

There are a number of different methods proposed in the literature to obtain desired networks, with one such model being a Markov Chain based model from Hui et al.<sup>4</sup>. It was used to analyze datasets that exhibited “trichotomy” (having several types of distributions within) and

perform better than approaches that relied on the data set only following one type of distribution, but it exhibited a higher level of complexity like many models commonly used in the field. Complex models are generally less understandable and can be more difficult to adequately manipulate, with the combination of those two factors resulting in a gap between application and research. Simpler models are more likely to be adopted by industry, easier to adjust and understand, and, as such, have greater capacity to create a meaningful contribution to the world.

The model proposed in this study will be based on the Molloy-Reed model with an algorithm inspired by Markov Chains and the model proposed by Hui. It will have an emphasis on simplicity while maintaining an adequate representation of road networks. The model will assume that the degree distribution of the network is known, and will use the Molloy-Reed model to create a preliminary microstate. Then the algorithm that will be proposed will attempt to increase the clustering coefficient using a greedy algorithm. Commonly used transportation networks will be used in the study<sup>5</sup>. They were chosen because they are properly labeled and small enough to run many tests on quickly.

## **BACKGROUND**

In order to effectively determine the feasibility of a simple Markov Chain-based model for road networks, we opted to examine a variety of papers regarding transportation networks and models that were based on Markov Chains. By doing so, we hoped to better understand possible approaches, their benefits and drawbacks, and key features to look out for when designing our model and its algorithm, even though we didn't directly use Markov Chains.

The work of Barber et al. presents two optimization models used for generalized transportation networks, focusing on minimizing objectives such as production, costs, and capital. While the work examines more specific data about the nodes within the network they're examining, the general approach laid out provides valuable insights on approaching a network problem, especially a transportation network since it shares similarities to the road networks this paper is working with. Both models incrementally improve the network by altering the network. Barber et al. attempts to choose incremental improvements based solely on the direct effects on costs. This idea of incremental improvements is referred to in this paper as "greedy" algorithm. Unlike Barber, however, the object of the incremental improvements will be the clustering

coefficient of the network. As such, this paper provides several key tips on how to approach designing a network problem for our work.

One Markov Chain-based approach that we found came from Hui et al. who utilized Markov Chains to create a unified framework that describes complex networks' evolution processes. The paper addressed the issue of node-degree distributions in complex network models where models employ only one type of distribution despite data sets exhibiting what was call a "trichotomy," containing power-law distributions in the middle section of the data distribution and exponential-alike distributions on both sides. The framework leverages Markov Chains to better model this trichotomy by splitting the process of changing a node's degree into three phases: initializing, fast-evolving, and maturing; this serves to generalize traits of many existing models in literature and introduce extra features existing complex network models lacked. The paper's scope is wider than what our proposed work aims to accomplish. Hui et al. leverage Markov Chains to better model complex networks, whereas our project aims to model road networks with a focus on simplicity. However, the work of Hui et al. is helpful as reference for thinking through how to best formulate and integrate network evolution into our model.

In Ghoshal, G., Chi, L. & Barabási's paper on elementary processes in network evolution, the addition of internal link and the removal of links, also known as "edge deletion", is discussed with relation to a theorized degree distribution. The paper calculates the theoretical degree distributions for different simulation parameters, like the likelihood of edge, link, and node deletions and additions, and compares them with the simulations. The paper's results discuss that it is difficult to use evolving networks to achieve a desired degree distribution, and that there is a complex interplay between the parameters that influence the resulting distribution. However, the paper indicates that it is possible to steer the model to a desired distribution. For instance, allowing new links to be established between preexisting nodes, with a bias, increases the degree of nodes with an already high degree. The framework in the paper is a good background for understanding the difficulty in maintaining the initial degree distribution that will be generated for our proposed model and thinking about ways to overcome this issue.

There are a number of routes we could've taken when designing and implementing our network. The above works are more complex in their scope and networks than the proposed model, but they provide ample information on approaching a network problem from different

perspectives while showing different ways to adjust and implement models to address key features or generate interesting insights. This material can present challenges to readers unfamiliar with the topic due to the complexity of languages and ideas presented, so we sought to use their concepts in order to create more simple, easy-to-understand models in our work.

## METHODOLOGY AND RESULTS

To begin the process of developing a simple and easy to understand model, it is desirable to create a schema which will allow for consistent testing and validation throughout iterations.

Initially, we attempted to correlate network parameters like clustering coefficient, associative coefficient, and diameter with the number of nodes in the network using networks and data we acquired from the Transportation Networks data repository<sup>5</sup>. The idea was to allow for a wider, more consistent set of testing data that, which we believed would show the large possible applicability of the model. The model would be fed realistic parameters for number of nodes, and clustering coefficient and validated against the diameter and associative coefficient that correlate to the number of nodes. The preliminary data analysis is shown in Figures 1-3 for the selected networks and data in Table 1.

Network	Zones	Links/Edges	Nodes/Vertices
Anaheim	38	914	416
Barcelona	110	2522	1020
Berlin-Friedrichshain	23	523	224
Berlin-Mitte-Center	36	871	398
Eastern-Massachusetts	74	258	74
Sioux Falls	24	76	24
Sydney	3264	75379	33837
Winnipeg	147	2836	1052

*Table 1: The selected networks and their number of links and nodes*

In Table 1, we can see the variety of networks chosen for our work; we made sure to pick networks with different characteristics in order to more thoroughly test our models.

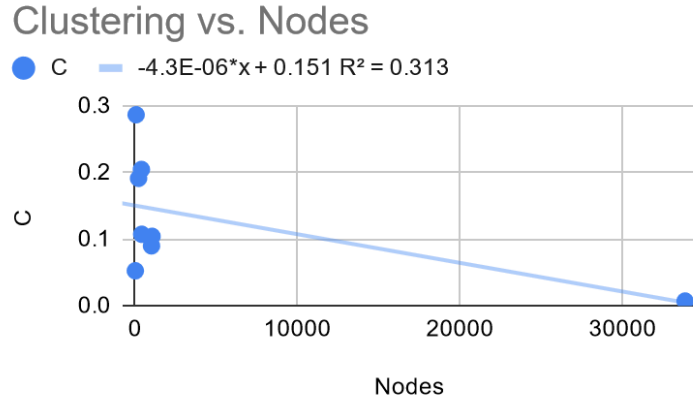


Figure 1: Clustering coefficient versus node size for networks in table 1 with best fit line.

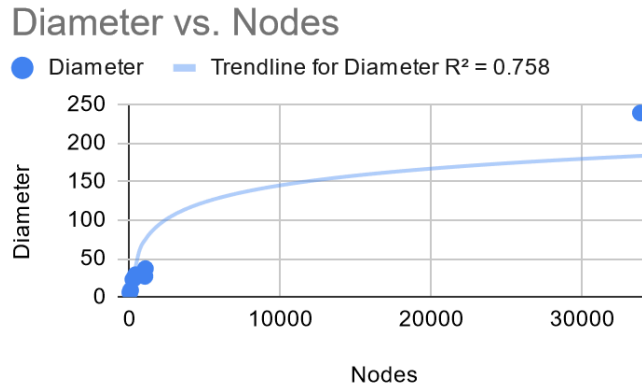


Figure 2: Diameter versus node size for networks in table 1 with best fit line.

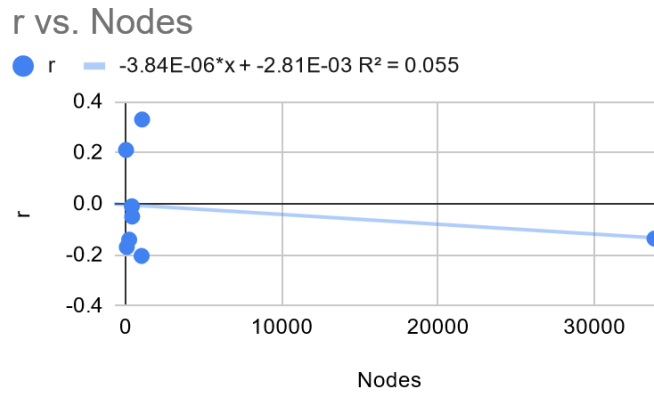


Figure 3: Associative coefficient versus node size for networks in table 1 with best fit line.

During our analysis in Figures 1-3, it quickly became apparent that an attempt to find simple parametric models of road networks would fail. The  $R^2$  values of the best fit lines for the clustering and associative coefficients indicated that there was little correlation between the number of nodes and the value of the coefficient. It was discovered that the diameter and the node size follow a logarithmic relationship which roughly follows the approximation of expected diameter for a large sparse random graphs<sup>9</sup>:

$$D \approx \frac{\log n}{\log \langle k \rangle}$$

The preliminary analysis showed that it would be difficult to correlate “realistic” network parameters from the number of nodes. Therefore, it was decided that the model would be given the degree distributions, and clustering coefficient of the real networks and be validated against the real values from those networks. The drawback of this is that the proposed model can only attempt to recreate already existing networks with known clustering coefficients and degree distributions.

The validation of the model was originally defined using three categories:

1. Did the model achieve the desired clustering coefficient?
2. Did the model maintain the desired degree distribution?
3. Do other network parameters, like the associative coefficient, of the generated model match the real network?

For this study, the clustering coefficient was the primary parameter that the model would be focused on, but any network parameter could be used with greedy algorithms. The decision to use clustering was because clustering proved to be consistently higher than what a random network would generate resulting in models that only need to increase the coefficient, and not decrease it. Next, the desire to maintain the degree distribution came from real life road networks. Most road networks have fairly small degrees since in reality the nodes are intersections or points where the speed of the road changes. The goal of this proposal was to model road networks and having degree distributions that contained a large number of nodes with high degrees was not realistic, hence the second validation. The third validation was a check against the very likely possibility that the algorithm was only improving the clustering coefficient and not making the networks more similar to road networks. By comparing the

networks' other parameters that were not actually improved throughout the model, it is possible to see if we are creating road networks or just networks with a desired clustering coefficient.

## **MODELS**

The models proposed in the literature are greedy. Initially, a random Molloy Reed network is created that follows the desired degree distribution. This initialization allows the algorithm to start with the desired degree distribution. While this does assume that the degree distribution is known it allows the study to focus on fitting clustering coefficients well rather than attempting to achieve two targets at the same time. Next, the random networks clustering coefficient is calculated. If it is not within the allowable threshold, the algorithm continues by going to each node and attempting to find links that improve the clustering coefficient. For each link established a link is also removed. Each version of the algorithm uses different strategies to find the pool of candidate nodes with which each node can link to. Finally, when the clustering coefficient is within threshold, the algorithm stops and the unconnected components are artificially connected to the largest component. This final step is done to more properly model real road networks where unconnected components do not occur.

### **MODEL 1**

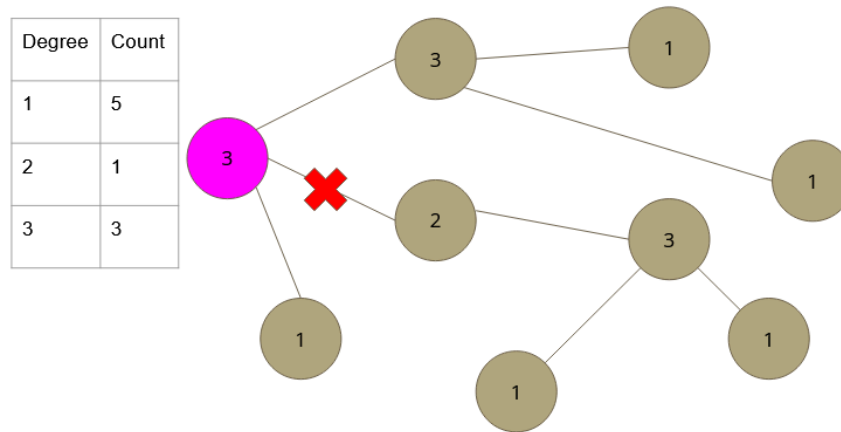
The first model's strategy of finding candidate links was entirely random. The model would require the initial degree distribution, the desired clustering coefficient, and a number, which was much smaller than the size of the network. Then, the algorithm will iterate through all the nodes until the stopping criteria is met. For each target node, a random link from the target node would be removed, and the preset number of nodes would be selected as potential nodes. Finally, the algorithm would find the link that increases the clustering coefficient the most from the potential node pool and establish that link. The algorithm proceeds to move on to the next node and repeat this process until it stops.

Model 1 Algorithm (Figures 4-6):

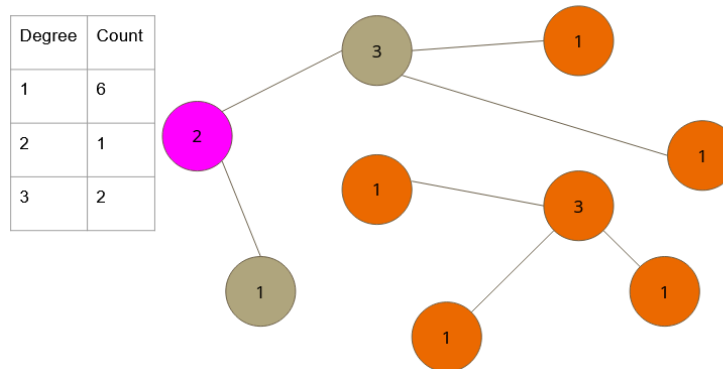
1. A Molloy Reed network is created
2. Network parameter (clustering coefficient) is calculated



3. For each node:
  - a. A random link will be chosen to be removed (that is not the new link) to maintain the degree distribution
  - b. A number  $X \ll n$  nodes will be selected as potential nodes to link to
  - c. Each possible link from the pool of potential nodes will be evaluated on which one most improves the clustering coefficient
  - d. The best link will be chosen
4. Steps 3 will be repeated until a stopping criteria is met (clustering coefficient is within acceptable bounds)
5. Unconnected parts will be artificially connected to the network



*Figure 4: Model 1, Step 3a - Random Link Removal*



*Figure 5: Model 1, Step 3b and 3c - Potential Nodes are chosen and analyzed*

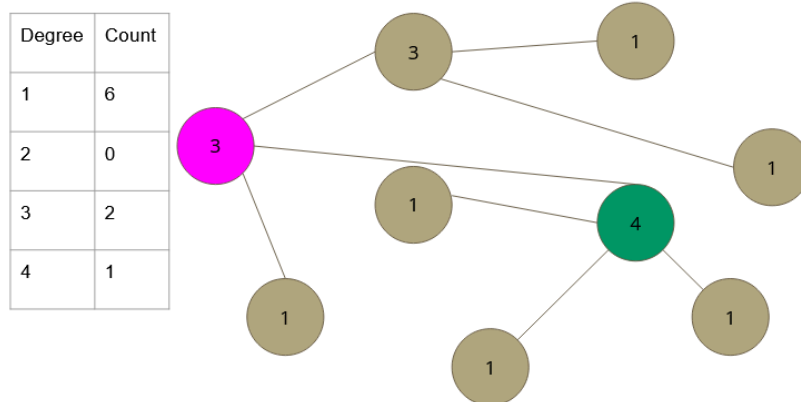


Figure 6: Model 1, Step 3d - The best link is chosen

## MODEL 1 RESULTS

The model did approach the desired clustering coefficients, but it failed in maintaining the degree distribution and in attaining a reasonable associative coefficient, as shown in Figure 7 and Table 2. While we expect to see better performance in terms of approaching coefficients with more run time for the model, it still presents concerns about Model 1's design and behavior.

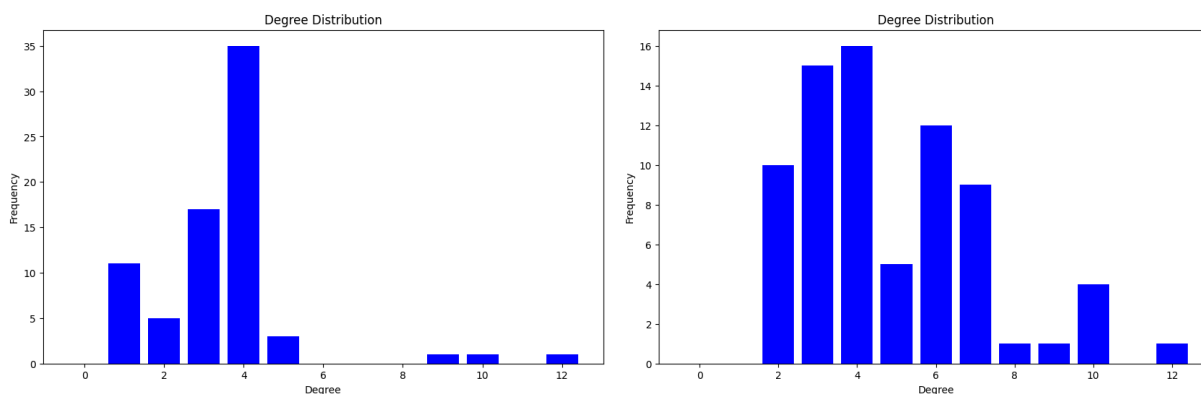


Figure 7: Degree Distribution Comparison for the Original Eastern Massachusetts network (left) and Model 1 (right)

A brief glance here at Figure 7's comparison of the degree distribution for the original Eastern Massachusetts network and the degree distribution for the network generated by Model 1 based on Eastern Massachusetts showcases the extreme change our model brought about; instead

of maintaining a similar degree distribution, we see a right-skew towards a more normalized degree distribution, with nodes clustering around degrees 5 and 6 rather than exhibiting the large quantity of nodes with degree 4 as in the original. Additionally, we note the increase in nodes with higher degrees, indicating that Model 1's behavior favors degrees of larger size.

CITY	Diam. Orig.	Diam. New	C Orig.	C New	r Orig.	r New
Anaheim	26	7	0.1076	0.0511	-0.0495	0.0332
Barcelona	27	9*	0.0902	0.0226	-0.2036	-0.0516
Berlin-Friedrichshain	23	6	0.1915	0.1184	-0.1398	0.0157
Berlin-Mitte-Center	29	6	0.2048	0.0604	-0.0092	-0.0294
Eastern Massachusetts	9	7	0.2869	0.2615	-0.1687	-0.0436
Sioux Falls	6	3	0.0528	0.7716	0.2112	-0.3782
Sydney	239	34*	0.0074	0.0002	-0.1357	0.0087
Winnipeg	37	9*	0.1045	0.0242	0.3304	-0.0102

*Table 2: Result comparison between Original Network and Model 1 trained for 60 minutes*

Further examining the output of Model 1 in Table 2 across all of our networks, we see a varied performance. One of the biggest trends that stands out here is the overall sharp decrease in the diameter of the networks, with some of them being unconnected (represented with a \* on their diameter). This suggests a higher interconnectedness of the network overall compared to the original, but we understand that this was an inherent property of Model 1's network generation given the friendship paradox: a node's neighbors likely have more neighbors than the node itself, meaning that a random reassignment of links will tend to result in new links to nodes with higher degree distributions, resulting in an overall decrease in the network diameter. As for the other metrics, we see the clustering coefficient approach that of the original network along with improved assortativity coefficients for networks such as Eastern Massachusetts, Anaheim, and Berlin-Friedrichshain, which is a big positive. We even note a particularly great performance on the Sioux Falls network, achieving a clustering coefficient of 0.7716. However, we also note that

this performance does not extend to all networks; many of Model 1's performances on networks here have clustering coefficients close to 0; this could be due to a multitude of factors, but we believe that this is largely due to the algorithm of Model 1 not lending itself to shorter runtimes. With longer runtimes beyond the 60 minute limit we imposed, we hypothesize that we would see much better results across networks that are more in line with the results shown on the Eastern Massachusetts network. However, even if we achieve better results for the clustering and assortativity coefficients, Model 1 still fails to maintain the degree distribution of the original network, meaning that while it can serve as a solid baseline for complex models, it doesn't fully embody the characteristics of the original network examined. As such, we returned to the drawing board after achieving these results to see how we can iterate on Model 1 to achieve our initial goals.


## **MODEL 2**

Model 2 was created in response to the issue Model 1 had with achieving clustering and assortativity coefficients in the generated models similar to those of the original network. In order to increase the clustering coefficient, we sought to create triades within the network with our algorithm, so Model 2 limits the pool of nodes that are evaluated at each step to nodes that are neighbors of neighboring nodes. However, this method has its own major drawbacks. First, it does not address the limitations of Model 1 regarding degree distribution like Model 3 (a separate approach discussed below) does. Second, by limiting the pool of potential nodes so drastically, it is possible that the algorithm will create highly clustered clumps that are not connected, rather than create a large connected component. Third, due to the additional computational complexity of finding neighbors of neighbors it is expected that this algorithm will take much longer to complete compared to Model 1, meaning that with a given finite limit, we expect to see worse performance for Model 2 compared to Model 1 and to the original networks due to requiring more time for meaningful changes to occur.

Model 2 Algorithm (Figures 8-10):

1. A Molloy Reed network is created
2. Network parameter (clustering coefficient) is calculated

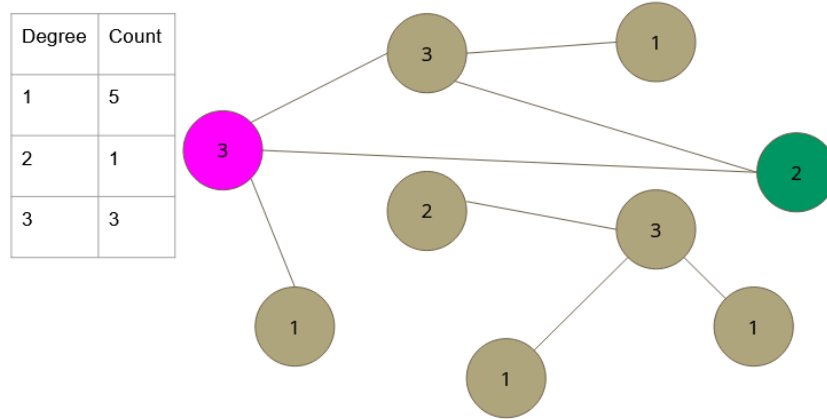
- | Degree | Count |
|--------|-------|
| 1      | 5     |
| 2      | 1     |
| 3      | 3     |



The graph consists of 10 nodes and 10 edges. The nodes are colored based on their degree: magenta for degree 2, olive for degrees 1, 2, and 3, and orange for degree 1. The graph is connected and contains a cycle of length 4.

Degree	Count
1	6
2	1
3	2

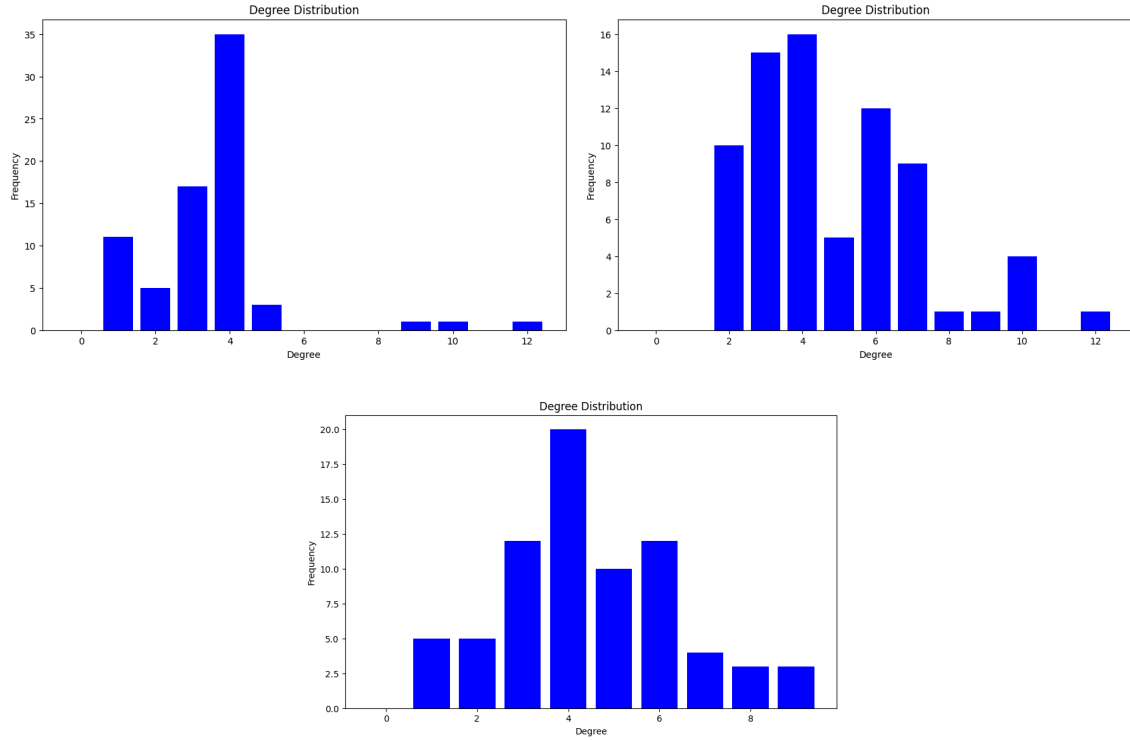
13



*Figure 10: Model 2, Step 3d - Connect with the best node*

## MODEL 2 RESULTS

Much like with Model 1, we saw the model fail to meet expectations; however, Model 2 did so in almost an inverse manner compared to Model 1 as can be seen below. The model did not get close to approaching the desired clustering and assortativity coefficients as we initially planned, but it demonstrated much greater aptitude in maintaining the degree distribution of the original network. Should the model be given much longer runtimes, we believe that we could see a great improvement in performance, but these runtimes would likely be much longer than that of Model 1 due to the algorithmic complexity differences, so the end use cases for Model 2 would differ.



*Figure 11: Degree Distribution Comparison for the Original Eastern Massachusetts network (top left), Model 1 (top right), and Model 2 (bottom)*

In Figure 11, we compare the resulting degree distributions from Model 1 and Model 2 to that of the original model. One of the biggest issues with Model 1 was that its resulting degree distribution exhibited right-skewed normalization compared to the original, limiting how useful it would be to analyze that network's node layout in order to understand the original network better. Model 2's algorithm, while not initially designed to do so explicitly, created a degree distribution more closely resembling that of the original network. While there is still some normalization of the distribution occurring that expectedly results in quantities of nodes with degree distributions not seen in the original (as can be seen in the appearance of nodes with degrees 6 through 8), we do see that nodes cluster around degree 4 much more like they do in the original degree distribution. Model 2's degree distributions overall look more normalized, as opposed to the right-skewed distributions from Model 1, and more closely resemble the original degree distributions of the networks, indicating that Model 2's network characteristics are more strongly correlated to those of the original networks.

CITY	Diam. Orig.	Diam. New	C Orig.	C New	r Orig.	r New
Anaheim	26	6	0.1076	0.0	-0.0495	-0.0012
Barcelona	27	9*	0.0902	0.0	-0.2036	0.0020
Berlin-Friedrichshain	23	6	0.1915	0.0	-0.1398	-0.0328
Berlin-Mitte-Center	29	6	0.2048	0.0	-0.0092	-0.0106
Eastern Massachusetts	9	6	0.2869	0.0	-0.1687	-0.1408
Sioux Falls	6	7	0.0528	0.0528	0.2112	0.1669
Sydney	239	34*	0.0074	0.0	-0.1357	-0.0096
Winnipeg	37	8*	0.1045	0.0	0.3304	-0.0143

*Table 3: Result comparison between Original Network and Model 2 trained for 60 minutes*

Before moving on to the rest of the discussion of the results in Table 3, we once again note the expected behavior of the decrease in network diameter compared to the original network; in this case, it was even more likely to occur due to the search space consisting of neighbors of neighbors of a node, meaning that the friendship paradox was much more likely to be seen to a higher degree as compared to Model 1. Curiously, we find that some of the networks were still split into multiple components despite the algorithm being designed to ensure all nodes are on the same component; this is a route for future improvement as it suggests a redesign of how we're handling nodes to examine and where we assign them. We note that Model 2 has similar diameters to those of Model 1, but as the input network size grows, we expect to see Model 2 have smaller diameters, on average, compared to Model 1. As suspected due to the newly introduced algorithmic complexity and search space, Model 2 failed to achieve meaningful values for the clustering and assortativity coefficients across all networks within the allotted time limit of 60 minutes for each run, often finding its values at or very close to 0. Despite this seeming failure, we note the exception of its performance on Sioux Falls to those findings, where Model 2 not only achieved the desired clustering coefficient but did so while maintaining a good assortativity coefficient value that approaches the one from the original



Sioux Falls network. Given the much smaller size of Sioux Falls compared to the other networks we tested on in tandem with the lengthier runtimes we expected to need for Model 2, these findings give hope that Model 2 would demonstrate similar success in achieving closer clustering and assortativity coefficient values to those of the original networks while still maintaining a similar degree distribution, making the networks of Model 2 suitable for studies due to the characteristic similarity it'd show to the original networks.

### **MODEL 3**

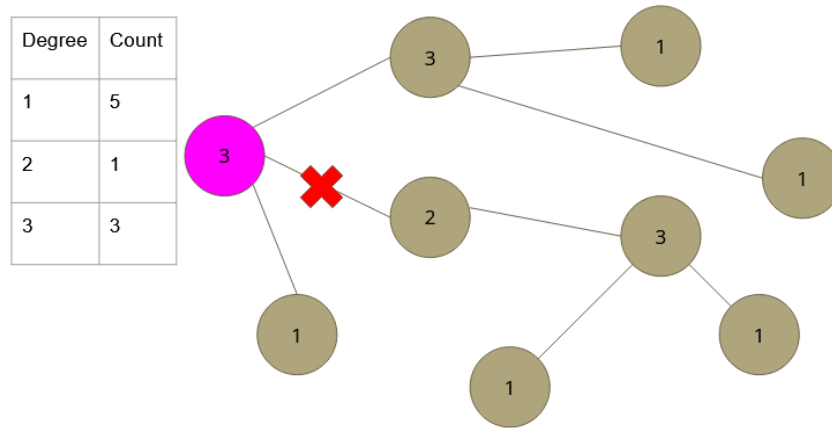
The third iteration in our work attempted to more directly address the changing degree distribution of Model 1. The degree distribution in the first model tended to “flatten out;” the number of nodes with degree 4 may be 35 in the original network and then be 16 in Model 1's version of the network. This occurred because while the node that was finding new links maintained its degree by removing a node and adding a node, the node that had a link removed had its degree reduced and the node with the new link had its degree increased. Taking this over a large number of nodes would lead to that normalization of the degree distribution that we saw in Model 1's results, with the right-skew towards nodes of higher degree because of the friendship paradox we've discussed so far.

To address this issue, Model 3 used a different strategy when choosing the potential link pool for each node: first, for a node, a random link will be removed, and the degree of the node that lost the link will have been found prior to the link's removal. Next, a set of potential nodes with the same degree as the node that lost the link will be proposed, and the potential link that increases the clustering coefficient the most will be selected for addition. By attempting to add links to nodes of a degree one lower than the degree of the node whose link was removed, we hope to approximately maintain the overall degree distribution of the original network.

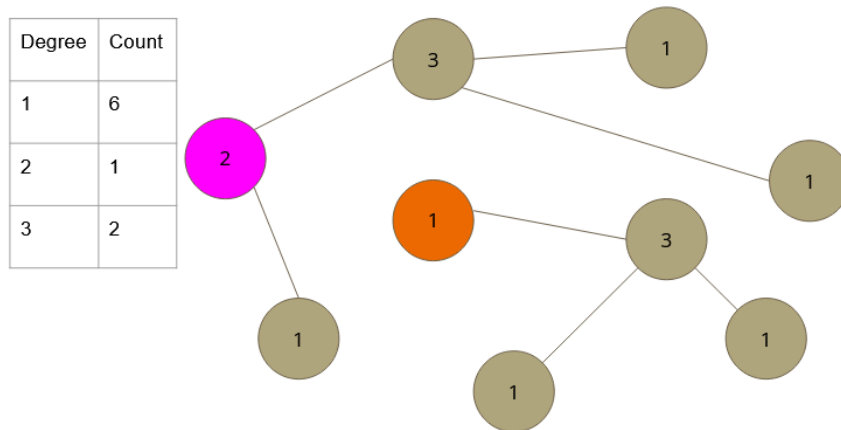
Model 3 Algorithm (Figures 12-15):

1. A Molloy Reed network is created
2. Network parameter (clustering coefficient) is calculated
3. For each node:
  - a. A random link will be chosen to be removed, node Z

- b. The degree of the node Z is found
  - c. X nodes with degree of node Z will be selected as potential nodes to link to
  - d. Each link will be evaluated on which one most improves the clustering coefficient
  - e. The best link will be chosen
4. Step 3 will be repeated until a stopping criteria is met (clustering coefficient is within acceptable bounds)
5. Unconnected parts will artificially connected to the largest component



*Figure 12: Model 3, Step 3a - Random Link Removal*



*Figure 13: Model 3, Step 3b - Find the degree of the node whose link was removed*

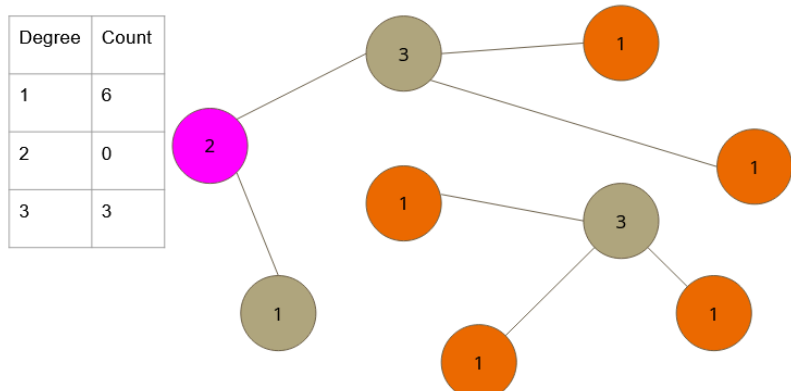


Figure 14: Model 3, Step 3c and 3d - Consider and evaluate nodes with that same degree

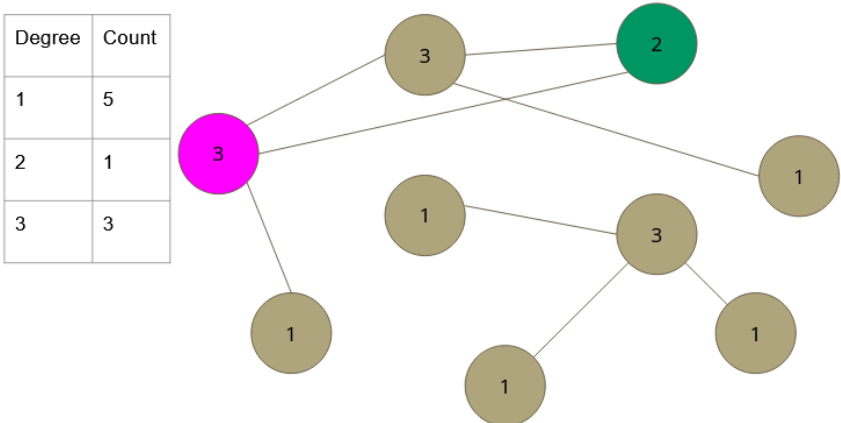
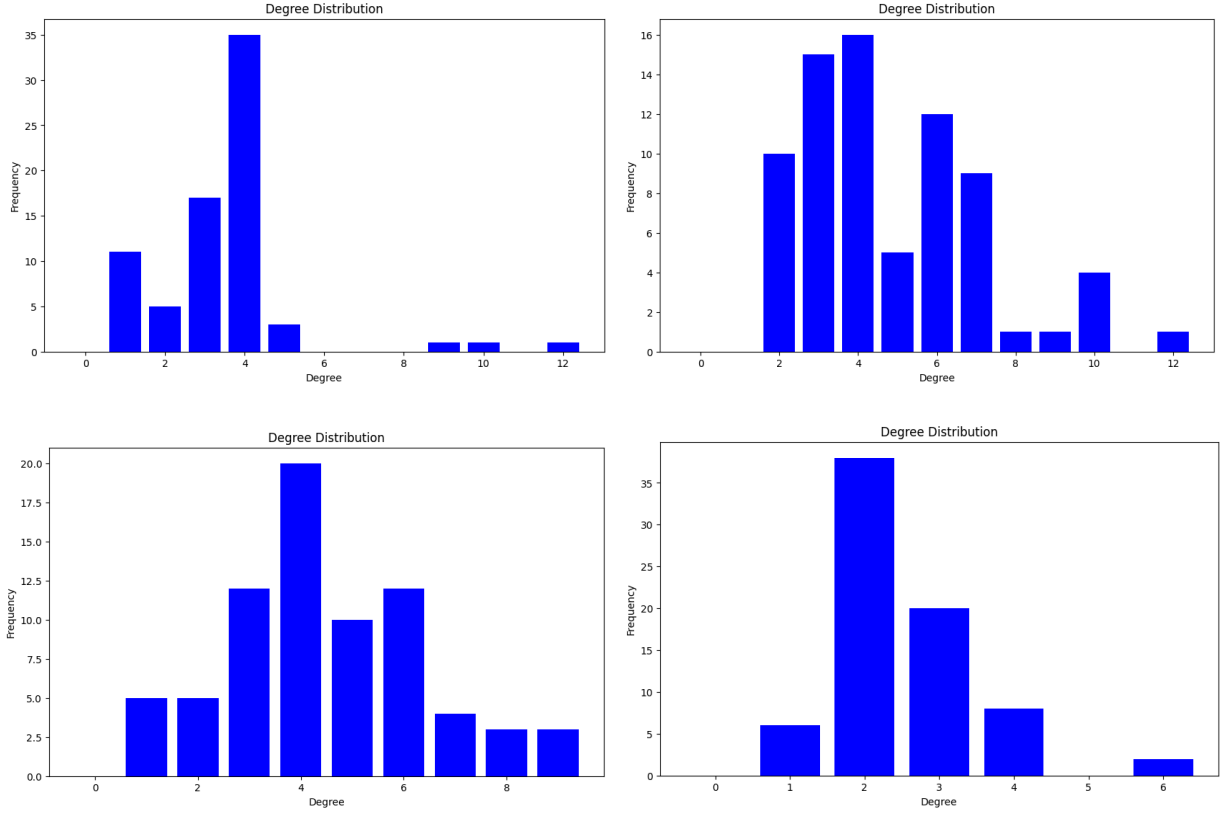


Figure 15: Model 3, Step 3e - Choose the best link

### MODEL 3 RESULTS

Whereas Model 2 sought to address the issues Model 1 had with approaching the original clustering and assortativity coefficients, Model 3 was made in order to address Model 1’s shortcomings regarding maintaining a degree distribution more similar to that of the original graph. Model 3, in turn, showed itself to be somewhat successful at doing so while matching or improving upon Model 1’s performance regarding the clustering and assortativity coefficients, which was phenomenal. We’ll further examine this in Figure 16 and Table 4 below.



*Figure 16: Degree Distribution Comparison for the Original Eastern Massachusetts network (top left), Model 1 (top right), Model 2 (bottom left), and Model 3 (bottom right)*

Being designed specifically for the purpose of maintaining the original degree distribution, Model 3 struggled somewhat more in this department than we had initially expected. As seen in Figure 16, Model 3 exhibits a right-skewed distribution like the other Models (albeit with slightly less normalization), but the majority of values can actually be found at a degree even lower than that of the original degree distribution, indicating that the algorithm potentially caused nodes to become linked to nodes with degrees lower than intended. This behavior extends to the other tested models (results available in our GitHub repository), with the run on Barcelona also finding itself largely centered around degree 2 while the run on Sioux Falls maintains a degree distribution roughly similar to its original network's. This indicates that the current version of Model 3 likely would left-shift the degree distribution of larger networks, but that can be addressed in future iterations; the current Model 3 performed well in this regard.

CITY	Diam. Orig.	Diam. New	C Orig.	C New	r Orig.	r New
Barcelona	27	23	0.0902	0.0019	-0.2036	0.0342
Berlin-Friedrichshain	23	17	0.1915	0.0360	-0.1398	-0.0724
Berlin-Mitte-Center	29	21*	0.2048	0.0254	-0.0092	0.0163
Eastern Massachusetts	9	15*	0.2869	0.2820	-0.1687	-0.0343
Sioux Falls	6	10	0.0528	0.0500	0.2112	0.1515

*Table 4: Result comparison between Original Network and Model 3 trained for 60 minutes*

As seen in Model 2, we still find some of the networks being split into multiple components in Table 4 despite the algorithm being designed to ensure all nodes are on the same component. Oddly, we also note an increase in the diameter for some networks such as Eastern Massachusetts and Sioux Falls, but this is likely due to the design philosophy of the algorithm preemptively addressing the friendship paradox we’d seen in the prior Models; since the nodes aren’t necessarily connecting to nodes with higher degrees than themselves, ‘shortcuts’ across the network aren’t made as easily, and many shortcuts are actually being removed. These are potential points for improvement in future iterations of this work. Model 3, while attempting to maintain the original degree distribution of the network, shows itself to successfully reach similar clustering coefficients and similar to improved assortativity coefficients on the input networks. We note that Barcelona and Berlin-Friedrichshain had clustering coefficients close to 0 with improved assortativity coefficients, but we expect that the clustering coefficients would increase with longer runtimes given the large difference in size between those two networks and Eastern Massachusetts and Sioux Falls. Given the overall performance of Model 3, we believe its conceptual algorithm would do well in terms of creating networks with similar or improved characteristics compared to the original network used, so it serves as a potential foundation for future work to iterate on in order to create a better model that provides more utility when modeling networks. Even in its current state, Model 3 still did relatively well in terms of achieving good clustering and assortativity coefficients while trying to maintain the degree distribution of the original network.

## CONCLUSION AND FUTURE WORK

The three Models we tested in our project had varying degrees of success and failure. Model 1 was able to achieve similar clustering coefficients while improving the assortativity coefficient but failed to maintain a similar degree distribution to the original network. To address these issues, we iterated upon Model 1 to create Models 2 and 3. Model 2, designed to improve the clustering and assortativity coefficients through the use of triads, failed to show significant results on the larger networks within our allotted training time, but it did show promise with its good performance on the Sioux Falls network, indicating that Model 2 could exceed Model 1's performance should it be given ample time to run. More notably, it did a far better job of maintaining a more similar degree distribution to that of the original network, showing its potential as a useful generator for baseline networks to use in testing and comparisons. Model 3 was designed to maintain the original network's degree distribution, and it proved somewhat successful in doing so (albeit with a left-shift in the distribution). More importantly, Model 3 did so while achieving similar clustering coefficients and similar to improved assortativity coefficients to those of the original networks used, indicating that Model 3 managed to perform decently well at emulating the key characteristics of the original networks! Models 2 and 3 both could use some slight tweaks to address having multiple connected components per network, but their core algorithms seem to be solid and likely would provide better results if retooled and adjusted. Model 1 does good comparatively on shorter time scales but would likely perform worse than the others should the models all be given longer runtimes, and a model that incorporates the strategies of Models 2 and 3 could prove to be much more successful for generating a road network with a desired degree distribution and clustering coefficient while maintaining elements of simplicity in its design and implementation, providing an easy but solid introduction to this field of study; the results of such a model would then also serve as a good baseline for comparison of more complex models that researchers may develop in their work rather than having to rely on either no baseline model or a pure random baseline model. Better results could be obtained with these Models if they were given much longer periods of time to run (we typically limited each run to 60 minutes) given the large size of many networks and the search space of our algorithms as well as the algorithms' complexity. Additionally, we believe that more interesting trends could be found through the use of more networks, especially larger

ones, in tandem with longer runtimes; our research particularly examined transportation networks, but we feel that the models could provide interesting insights if applied to other fields of study, such as biological and social networks. Further improvements and optimizations to the code would also be likely to show improvements in the models performance, particularly if they were optimized to cut down on runtime. While there is still plenty of work that could be done to advance and improve upon our project, we feel that we created interesting and (conceptually) simple models that can serve as strong baselines as well as alternative methods for studying existing transportation networks.

### **GENERATIVE AI DISCLAIMER**

This project made use of ChatGPT for assistance in generating code as well as discussing and going through concepts. Additionally, it was used to aid in adding type hints as well as comments to the code. Links to ChatGPT logs can be found at the GitHub repository below. ChatGPT was not used in any of the actual writing of the paper.

### **REPOSITORY**

<https://github.com/meetpatel450/cee520-final-project/tree/main>

## REFERENCES

1. Erdős, P.; Rényi, A. (1959). "[On Random Graphs. I](#)" (PDF). *Publicationes Mathematicae*. **6** (3–4): 290–297. doi:[10.5486/PMD](#).
2. [Gilbert, E.N.](#) (1959). "[Random Graphs](#)". *Annals of Mathematical Statistics*. **30** (4): 1141–1144. doi:[10.1214/aoms/1177706098](#)
3. M. Molloy and B. Reed (1995). [A Critical Point for Random Graphs with a Given Degree Sequence](#). *Random Structures and Algorithms* 6 161-180 .
4. Zhao, F., Zeng, X., Liu, H., & He, W. (2017). The impact of the built environment on bicycle commuting: Evidence from Beijing. *Scientific Reports*, 7(1), 4887. <https://doi.org/10.1038/s41598-017-03613-z>
5. Stabler, B. (n.d.). *TransportationNetworks*. GitHub. Retrieved April 11, 2024, from <https://github.com/bstabler/TransportationNetworks>
6. Barber, G. M. (1975). A Mathematical Programming Approach to a Network Development Problem. *Economic Geography*, 51(2), 128–141. <https://doi.org/10.2307/143069>
7. Ghoshal, G., Chi, L. & Barabási, AL (2013). Uncovering the role of elementary processes in network evolution. *Sci Rep* 3, 2920. <https://doi.org/10.1038/srep02920>
8. Hui, D.S.W., Chen, YC., Zhang, G. et al. (2017) A Unified Framework for Complex Networks with Degree Trichotomy Based on Markov Chains. *Sci Rep* 7, 3723 (2017). <https://doi.org/10.1038/s41598-017-03613-z>
9. Hackl, Jürgen (2024). [Random Graphs and Small World Networks](#) (PDF). CEE520/COS520 Advanced Topics in Network Science Github. February 27, 2024. <https://github.com/cis-teaching/cee520-documents-all/blob/main/handouts/05-handout.pdf>



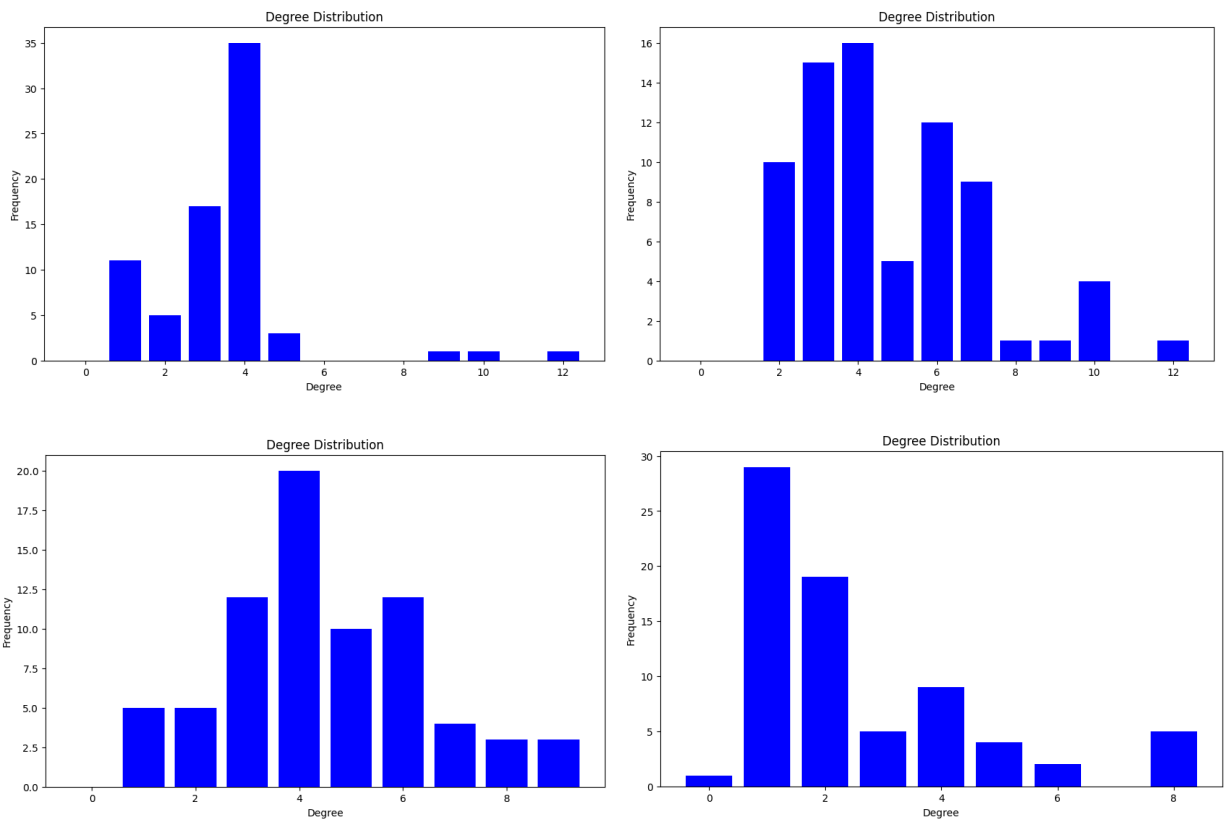
## APPENDIX

The following part of the paper is an alternative version, written based on results of a buggy version of Model 3 that used an incorrect value of the degree of the removed node (the degree tracked was 1 larger than it should've been). This section was included in the paper to showcase our iterative process and commitment to fully examining and understanding our models as well as to show the work we put into this project.

### ENDING OF THE PAPER WITH ORIGINAL (BUGGED) MODEL 3 RESULTS

#### MODEL 3 RESULTS

Whereas Model 2 sought to address the issues Model 1 had with approaching the original clustering and assortativity coefficients, Model 3 was made in order to address Model 1's shortcomings regarding maintaining a degree distribution similar to that of the original graph. However, we noticed that Model 3 struggled not only in regards to achieving the proper coefficients but also in attaining a similar degree distribution across all networks. In order to fully understand this, we'll examine the Figures and Tables below.



*Figure A: Degree Distribution Comparison for the Original Eastern Massachusetts network (top left), Model 1 (top right), Model 2 (bottom left), and Model 3 (bottom right)*

Oddly, despite being designed specifically for the purpose of maintaining the original degree distribution, we note that Model 3 struggled in this department. As seen in Figure A, Model 3 exhibits a right-skewed distribution like the other Models, but the majority of values can actually be found at a degree even lower than that of the original degree distribution. This indicates that the algorithm was flawed and caused the nodes to become linked to nodes with degrees lower than intended. While it seems to approximately work here for Eastern Massachusetts (and seems to perform better at this regard than the other Models), an examination of the rest of the graphs reveals the erroneous behavior of this implementation of Model 3 wherein it performs worse across the board compared to the prior two models.

CITY	Diam. Orig.	Diam. New	C Orig.	C New	r Orig.	r New
Anaheim	26	8*	0.1076	0.0512	-0.0495	-0.3949
Barcelona	27	8*	0.0902	0.0420	-0.2036	-0.4271
Berlin-Friedrichshain	23	9*	0.1915	0.0555	-0.1398	-0.3234
Berlin-Mitte-Center	29	8*	0.2048	0.0612	-0.0092	-0.4170
Eastern Massachusetts	9	8*	0.2869	0.1141	-0.1687	-0.2698
Sioux Falls	6	7*	0.0528	0.0500	0.2112	-0.0987
Sydney	239	29*	0.0074	0.0	-0.1357	-0.0124
Winnipeg	37	8*	0.1045	0.0406	0.3304	-0.4083

*Table B: Result comparison between Original Network and Model 3 trained for 60 minutes*

Table B validates the concerns we had over Model 3's performance; it was able to achieve weak (but positive) clustering coefficients for networks but had comparatively abysmal results for the assortativity coefficients, with some results hovering around -0.4! This informs us that this version of Model 3's outputs were abhorrent compared to the other Models and the original network, with characteristics that were incredibly dissimilar to those of the original network.

CITY	Diam. 60	Diam. 90	C 60	C 90	r 60	r 90
Anaheim	8*	9*	0.0512	0.0594	-0.3949	-0.4285
Barcelona	8*	8*	0.0420	0.0362	-0.4271	-0.4506
Berlin-Friedrichshain	9*	9*	0.0555	0.0702	-0.3234	-0.3874
Berlin-Mitte-Center	8*	9*	0.0612	0.0462	-0.4170	-0.3828
Eastern Massachusetts	8*	11*	0.1141	0.0741	-0.2698	-0.1483
Sioux Falls	7*	6*	0.0500	0.0514	-0.0987	-0.2260
Sydney	29*	30*	0.0	0.0002	-0.0124	-0.0184
Winnipeg	8*	8*	0.0406	0.0366	-0.4083	-0.4297

*Table C: Result comparison between Model 3 for 60 and 90 minutes*

We've already established that time was a constraining factor for running models, so perhaps running Model 3 longer would improve our results? Well, as we can see in Table C, that is not the case. Model 3 not only failed to significantly improve across most networks when given more time, it actually decreased in performance with most 90 minute results noting lower clustering and similar to worse assortativity coefficients compared to the 60 minute results. This indicates an overall failure in performance from Model 3 compared to our prior models. Despite this performance, we still believe that Model 3's conceptual algorithm would do stellar in terms of creating networks with similar or improved characteristics compared to the original network used, so Model 3 serves as a potential foundation for future work to iterate on in order to create a better model that provides more utility when modeling networks.

## CONCLUSION AND FUTURE WORK

The three Models we tested in our project had varying degrees of success and failure. Model 1 was able to achieve similar clustering coefficients while improving the assortativity coefficient but failed to maintain a similar degree distribution to the original network. To address these issues, we iterated upon Model 1 to create Models 2 and 3. Model 2, designed to improve the clustering and assortativity coefficients through the use of triads, failed to show significant results on the larger networks within our allotted training time, but it did show promise with its

good performance on the Sioux Falls network, indicating that Model 2 could exceed Model 1's performance should it be given ample time to run. More notably, it did a far better job of maintaining a more similar degree distribution to that of the original network, showing its potential as a useful generator for baseline networks to use in testing and comparisons. Model 3 was shown to be less than stellar across the board; designed to improve the similarity of the resulting degree distribution to the original network's, it failed to do so while also achieving far worse assortativity coefficients, and its results even seemed to suffer with more time allotted. However, its core algorithm concept seems to be solid and likely would provide better results if retooled and adjusted. Model 1 does much better comparatively on shorter time scales but would likely perform worse than the others should the models all be given longer runtimes. A model that incorporates the strategies of Models 2 and 3 could prove to be successful for generating a road network with a desired degree distribution and clustering coefficient while maintaining elements of simplicity in its design and implementation, providing an easy but solid introduction to this field of study; the results of such a model would then also serve as a good baseline for comparison of more complex models that researchers may develop in their work rather than having to rely on either no baseline model or a pure random baseline model. As for future work upon this project, we believe that better results could be obtained if the models were given much longer periods of time to run (we typically limited each run to 60 minutes) given the large size of many networks and the search space of our algorithms as well as the algorithms' complexity. Speaking of networks, we believe that more interesting trends could be found through the use of more networks, especially larger ones, in tandem with longer runtimes; our research particularly examined transportation networks, but we feel that the models could provide interesting insights if applied elsewhere such as biological networks. Additionally, further improvements and optimizations to the code would also be likely to show improvements in the models performance, particularly if they were optimized to cut down on runtime. While there is still plenty of work that could be done to advance and improve upon our project, we feel that we created interesting and (conceptually) simple models that can serve as interesting baselines or alternative methods for studying existing transportation networks.