Consumer-Resource Relationships in the Messel Shale Food Web Through the Probabilistic Niche Model

Evan Meany and Nate Christy
Computer Science, University of Colorado Boulder

I. Introduction

The Messel Shale Food Web is an anomaly within the space of Paleontology due to the impressive number of organisms preserved by the oil shale rock that comprises the site. It is the best fossil cite for understanding biology and ecology withing the Paleogene Era, which is important because this Era is categorized by the emergence of mammals [8]. This leaves this food web as one of, if not the best ancient food web that still has much in common with an extant (modern) food web. Therefore analysis of this food web can reveal insights about how food webs that contain similar species like birds and mammals have evolved over the last 50 million years.

Analysis of this specific network and food web has been done previously [6], but we wanted to take another look at it specifically using the Probabilistic Niche Model with Simulated Annealing. Exploring food webs using Niche models is a common idea for complex food webs, because it can accurately depict a network with only 3 main traits per species (niche position, diet position, and range of feeding). Using the Probabilistic Niche Model instead, we can also estimate maximum likelihoods of each species position in the Niche Model to get a more precise picture of the model [11]. With the likelihood variable, the certainty of the model's output is present, and this is valuable because with a 48 million year old web, there will certainly be many missing species which will decrease the accuracy of our model despite the Messel Shale food web being one of the most complete that exist from sites this old.

So, through this analysis of aspects of this food web network such as network analysis, trophic analysis, and most importantly the Probabilistic Niche Model, ideas such as ecosystem-stability, food web depth and complexity, and more can be extrapolated [9]. This exploration of ancient ecology will hopefully show information about the nature of producer-consumer relationships in the Eocene Era and how they have changed over time.

II. Methodology

The methods used in this project can be split into four parts: sectioning the data, performing simple network analysis, doing trophic level analysis, and creating a Probabilistic Niche Model (PNM).

A. Data

The data found on this network was provided in the form of an Excel workbook [5]. This workbook include spreadsheets that describes the following: node ID, higher-level group, node name, trophic role (1 and 2), habitat, certainty, lines of evidence, Messel links (between consumer and resource), and associated link certainty. For our purposes, we loaded the edge data into Python as a directed Networkx graph. The edge certainty was made to be the 'weight' of each edge, giving our graph a kind of color gradient giving preference to links that are more certain. Other important data to us included the trophic role and habitat of each node. This data was loaded into Python dictionaries.

This food web can actually be separated into two groups: the terrestrial group and the aquatic group. The former consists of nodes in habitat group 1 of which lived primarily on the land. The latter consists of habitat group 2 of which primarily lived in the lake. The last group (habitat group 3) was inserted into both networks and considered 'amphibious' for our purposes. Although it should be noted there is still exists predation overlap between species that fall into the purely terrestrial and purely aquatic habitat groups.

B. Network Analysis

After generating the network data into a usable form, we conducted simple network analysis on both of the networks. This included number of species, number of consumer-producer interactions (edges/links), mean degree, clustering coefficient, and mean geodesic distance. These were done using the Networkx built-in functions, as well as methods we learned in lecture and implemented on the early problem sets [2], [4].

C. Trophic Level

Initially, to get a better grasp of the hierarchy of the network we were working with, and to provide a baseline for the PNM, we did a basic trophic level study on the network. After researching various methods for determining the trophic level of a species, we decided upon the Chain-Averaged trophic level Algorithm [1],[10]:

$$TL_j = 1 + \frac{1}{n}\sum_{i=1}^{n} l_{ji} \tag{1}$$

The above calculates the trophic level of each node based upon its average distance to every Basal node it can reach. Basal nodes are the primary producers in the network and, as such, have no prey of their own. The total number of Basal species that any particular species j has connections to is represented by variable n. l_{ij} is the shortest path length from node j to Basal node i.

By performing the Chain-Averaged trophic level algorithm, we were able to store the trophic level of every species in a Python dictionary for each network. Next, we wanted to create a visualization of the network. To do this we first wanted to categorize the species in each network. Using the trophic role dictionary previously created, we condensed the roles of each species down from 16 to only 4. These four are plants, fungus/bacteria, invertebrates, and vertebrates. these were then colorized: green, blue, yellow, and orange respectively (As seen in figures 4 and 5 below). Next, the graph was created by making the z position of each node correspond with its calculated trophic level. The x and y positions of each node were randomized. After each node was plotted on the graph, we iterated through all the edges to connect the graph. Each edge was colored based upon its certainty: black, dim gray, and light gray.

D. Probabilistic Niche Model

[11],[12] The PNM uses a probability function to calculate the probability that a consumer i feeds on a resource j according to the following equation:

$$P(i, j|\theta) = \alpha \prod_{d=1}^{D} exp(-|\frac{n_{d,j} - c_{d,i}}{r_{d,i}/2}|^e) \tag{2}$$

In the above, $n_{d,j}$ is the niche position of potential resource j and $c_{i,j}$ is the feeding position of potential consumer i. Essentially, this equation is checking to see if the niche position of the resource is in range of the diet position of the consumer. The other parameters adjust the range around the diet position of node i. $r_{d,i}$ is the feeding range of i and e affects the cutoff rate of the function. The last parameter, α, is the probability that i eats j when j is located at $c_{d,i}$. Sp $P(i, j|\theta)$ is simply the probability that node i eats node j given the set of parameters θ. Also, in the model we implemented, we set D equal to one so that the niche space was single dimensional.

The next step was optimizing the set of parameters to produce an accurate set of probabilities. First, we had to determine which variables were to be optimized. Initially, we wanted to optimize for $n_{d,j}$, $c_{d,i}$, and $r_{d,i}$; however, we settled on fixing the niche position of every species initially. To set a baseline for order in the niche space, the space was initialized so the niche position of every node corresponded to its trophic level. In this way we were able to utilize the trophic level analysis done previously. The equation for the niche position then is as follows:

$$n_i = \frac{TL_i - TL_{min}}{TL_{max} - TL_{min}} \tag{3}$$

From this equation, we can see that the niche space ranges from 0 to 1, where higher values correspond to consumers that are potentially higher in the food chain. Additionally, the trophic levels of all the species in the network were very slightly randomized so there were no ties in the niche space (namely at the Basal level).

The other two values left in the probability equation then were α and e. The former was set to be 0.999 in accordance with studies done previously. The latter was set to 2. Below shows a quick visualization of the PNM.

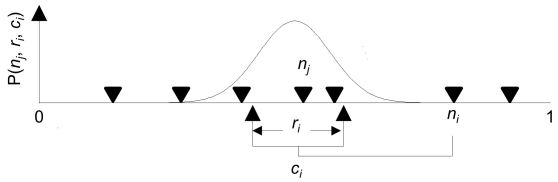

Fig. 1 Niche Space

All that was left now was to optimize the model's free parameters. First, $c_{d,i}$ and $r_{d,i}$ were initialized to random values between 0 and 1. However, the Basal nodes were initialized so that their diet positions were equal to 0 and their feeding range were very close to zero (10^{-6}). Other considerations were made for the specialists in the network (i.e. nodes with a single out-edge). The diet positions of these nodes were set equal to the niche position of their prey and their feeding ranges were set to a value very close to 0.

It was difficult to find specifics done in studies previously as to how a multi-variable optimization problem should be conducted for a network. We chose to go with a combination of simulated annealing that incorporated some aspects of the greedy heuristic used in class [3],[7]. Initially, the optimization calculates the likelihood function of the initialized network according to:

$$\sum_i \sum_j ln \begin{cases} P(i, j|\theta) & if X_{i,j} = 1 \\ 1 - P(i, j|\theta) & if X_{i,j} = 0 \end{cases} \tag{4}$$

$X_{i,j}$ is the Boolean existence of a connection between node i and j. It then loops through all the viable nodes in the network. For this model viable nodes were those that were not considered either Basal nodes or specialists. For each loop, the model calculates a Gaussian random number and applies it as a delta to the free parameter. It then recalculates the likelihood function. Next, it applies the Metropolis criterion which says the change made will be accepted according to:

$$u > \frac{L_{t+1}}{Lt} \tag{5}$$

The variable u is a uniformly random number between 0 and 1. L_{t+1} is the likelihood of the model with the change applied and L_t is the likelihood of the model with the previously accepted change. The model then does a check to see if the new likelihood is better than the best likelihood accepted thus far and if it is, it stores the new likelihood and the free parameters associated with it.

To incorporate both free parameters ($c_{d,i}$ and $r_{d,i}$), the model performs the loop outlined above for all the viable nodes in the network by only making changes to one parameter at a time. The Gaussian random change for the first loop had a standard deviation of 0.1 (where the parameter was $c_{d,i}$) and for the second loop it had a standard deviation of 0.2 ($r_{d,i}$).

Unfortunately, the model created was very computationally expensive as for each iteration of a single loop it had to calculate the likelihood of the entire network ($O(n^2)$). As such, we could not fully optimize the model and iterated the loops only a limited number of times. Additionally, due to the scale of the Terrestrial network, we only applied the model to the Aquatic network which was much smaller.

To determine the accuracy of the model we calculated the fraction of correctly predicted connections according to:

$$f_E = \frac{\sum_i \sum_j X_{i,j} P(i,j|\theta)}{E} \tag{6}$$

In the above equation, E is the total number of edges in the network. Additionally, we calculated the True Positive Rate and False Positive Rate of the predictions made relative to the edge list of the network.

III. Results

A. Network Analysis

	Terrestrial Network	Aquatic Network
Number of Species	594	49
Number of Nodes	691	120
Number of Edges	5440	444
Mean Degree	7.87	3.70
Clustering Coefficient	0.0388	0.1326
Mean Geodesic Distance	2.84	2.16

Table 1 Network Summary Statistics

This basic overview of the network can give suggestions both about the differences between both the two different habitats, and the food web as a whole. First we see that of the 700 total taxa, the terrestrial group has over 10x the species and links versus the aquatic group. The difference in nodes versus species is due to the aforementioned consumer-resource relationships that go between the two systems, but did not qualify the involved species as amphibious (habitat group 3). An example of this would be a bird hunting a fish.

Next the mean degree of the terrestrial system is over double that of the aquatic. This tells use there are significantly more inter-species relationships for the average land species, which is likely due just to the much larger number of nodes in the land system, but could also suggest that aquatic and terrestrial ecosystems inherently have different averages in the number of species consumed/consume them. The clustering coefficient of the aquatic group is much higher, although this is likely due to having far fewer species.

Finally the mean geodesic distance of the terrestrial group is around 1/3 greater than the aquatic, suggesting once again that the terrestrial system has more complicated relationships as well as a higher number of secondary and above level consumers. Although once again, it is plausible that the only factor involved in this is the disparity in number of species.

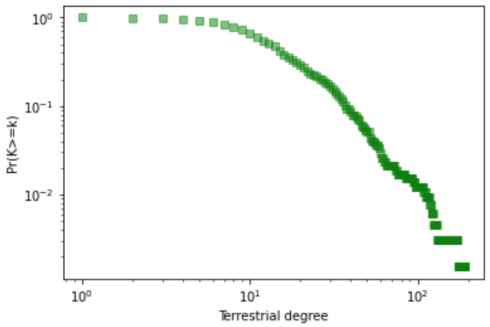

Fig. 2 Terrestrial Network CCDF

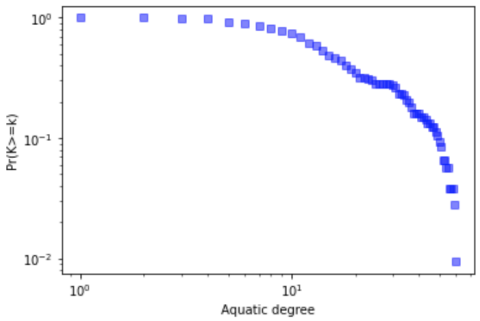

Fig. 3 Aquatic Network Trophic CCDF

Figures 2 and 3 above show loglog (with log_{10}) plots the Complementary Cumulative Distribution Functions of the two networks. In other words the log_{10} curve of probability that any one taxa has a number of consumer-resource (C-R) links. Note that the max degree of the terrestrial network is 162 (Pseudasturidae- an insectivorous bird), and

the max degree of the aquatic network is 60 (Tipulidae- a crane fly). That is a notable difference in maximums with a the terrestrial species max interacting with 23.4% of other species where the aquatic species max interacts with 50%.

From these plots we can make a few other observations. First, between the terrestrial and aquatic plots we can see that only about 10% of the species had over 16 C-R relationships, whereas for the aquatic network that 10% threshold is closer to 22. Also of interest is the sudden jump very near the bottom of the terrestrial network plot, which reveals that there were a very small number of taxa that had unusually many more C-R relationships.

B. Trophic Level Analysis

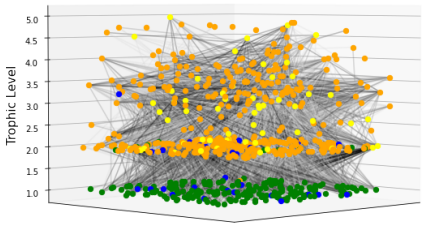

Fig. 4 Terrestrial Network Trophic Level Model

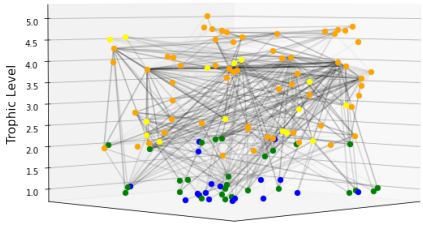

Fig. 5 Aquatic Network Trophic Level Model

Figures 4 and 5 above show the trophic level models that were created for the Terrestrial and Aquatic networks. The maximum trophic level of the Terrestrial network was found to be 5.545 and the minimum was 1 (the Basal nodes). The maximum trophic level of the Aquatic network was found to be 4.091 and the minimum was also 1. This checks out from what we noticed with Mean Geodesic Distance, because a higher max trophic level (and MGD) indicate a deeper food web, i.e. on average more species between the top consumers and the basal species.

To remind: *Blue* nodes are fungus/bacteria, *Green* nodes are plants, *Yellow* nodes are invertebrates, and *Orange* nodes are vertebrates.

From the figures we can already glean significant information into the systems and the data itself. From Figure 5 in particular we notice there is a distinct lack of basal (mostly blue/green) nodes in the aquatic network. This indicates the data set is missing many species because the invertebrates and basal species should be "hugely dominant" in a food web [9]. The terrestrial network seems to be a lot more evenly spread, although it still seems to have an unbalanced number of vertebrates. The likely explanation is again, that invertebrates, plants, and microbes are less likely to show up in fossil records. It is a small possibility on the other hand, that this is just the nature of this particular ecosystem.

Both of the figures show that there is a correlation between the trophic role of a species and its trophic level. At the lowest trophic level, most of the nodes are green nodes and there are some blue nodes. The middle of the network shows a mixture of mostly yellow nodes along with blue and orange nodes. The top of the network mostly has orange nodes. This generally matches what would be expected in terms of the distribution of trophic roles amongst trophic levels.

C. Probabilistic Niche Model

Number of Iterations	f_E	AUC
300	0.404	0.825

Table 2 PNM Performed on Aquatic Network

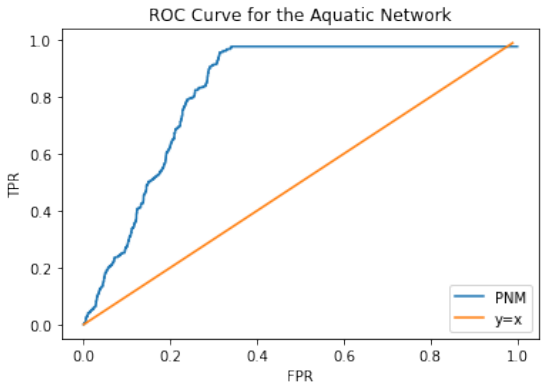

Fig. 6 ROC Curve

Table 2 and Figure 6 depict the accuracy of the PNM used on the Aquatic Network. In total, 300 iterations of the simulated annealing model were conducted. Optimization was not reached in this time frame; however, due to the large time complexity of the calculations performed, we decided 300 iterations were enough to show the results of the model. The f_E was found to be 0.404. The AUC was calculated as 0.825. The ROC curve above compares the PNM to a baseline model. From a first glance at the curve we can already see that the model did significantly better than the baseline prediction.

Overall the PNM model accurately predicted about 40% of the links correctly, and based on the AUC of 0.825 and f_E of 0.404 we know that the model was provably more accurate than any random link predictions. Note that this was again only for the aquatic network, as we did not reach an optimizations that allowed us to do a full simulated annealing run on the terrestrial network. It is easily plausible that our results would be worse for the terrestrial network, because that network has significantly more nodes and thus more potential false positives.

IV. Discussion

As stated previously, the PNM analysis performed was done only on the Aquatic part of the Messel Shale Food Web. Again, this is due to the scale of the Terrestrial network and the vast amount of time it would take to perform this type of analysis on it. If we had foreseen this issue, we perhaps would have chosen a different food web to study as just the Aquatic part of the food web is somewhat limited in terms of total interactions relative to the number of nodes. Despite this, we believe that the research we have done and computations we performed were successful in producing a good model under the circumstances as well as one that may be entirely unique it its implementation of simulated annealing.

From equation 6, we found that the model accurately predicted around 40 percent of the connections present in the network. This accuracy was at first disappointing, but compared to the existing study on this network by Dunne, Labandeira, and Williams, who found a PNM which predicted links accurately 26.6% of the time (on the terrestrial network), we think our model was actually reasonably successful. Dunne's study also got an accuracy of 50.2% when using extant (modern) food webs using their same implementation of the PNM [6]. That being said, because our implementation of the PNM was not very well optimized, resulting in barely getting 1 iteration to run after over an hour on the Terrestrial network, we cannot draw too many conclusions yet. It is more than plausible that the much higher number of nodes with equally low connectivity and clustering would result in a much lower accuracy if we could finish an entire run on the terrestrial system.

That level of accuracy on an extant food web is concerning for the PNM as a whole, because with extant food web we are much more certain that we are not missing any significant numbers of nodes. With a network constructed from a fossil record, it is more probable that links would be correctly predicted less frequently due to the whole ecosystem missing taxa that participate in the balance that the PNM is based off of. Once again going back to Dunne's study [6], we see that they attribute this low accuracy of the PNM to a network with a very high number of nodes and low connectivity. From our research and implementation of the PNM, this seems like a reasonable explanation. To create our model, we had to make educated inferences and calculations to get the Niche space position, diet position, and especially feeding range. The true data for each of these three factors is lost to time. Therefore the model itself relies on more guesswork than we would like, and since there is already a very low chance any one node n_i would connect to another other node n_j (691 nodes/5440 edges or 120 nodes with 440 edges), we have a significant room for error.

However, we did find that the model was at least somewhat accurate according to the ROC curve. We found that the AUC was 0.825 which means the model was more accurate than purely random connection chooses. This can also be seen in the depiction of the curve itself (Figure 6) as it is located to the left of the y=x baseline curve. Perhaps the reason for the somewhat high AUC value is the assumptions we made including the diet positions and feeding ranges of Basal nodes and specialists. These assumptions definitely bolstered the accuracy of the model. It is also notable that our result was over 300 iterations of simulated annealing (due to the immense time per iteration), so with more iterations or slightly tweaking the variables we may have been able to reach markedly better results. It should once again be briefly noted that these results were only for the smaller (120 nodes, 440 edges) aquatic network, which is likely to produce better results than the terrestrial.

The methods performed when creating this model perhaps make it unique in its implementation of simulated annealing. When researching methods, there was an unfortunate lack of scientific studies that outlined the specifics in performing a PNM analysis on a food web. We then decided to take pieces from multiple studies. For one, we decided to initialize the niche space in a similar manner to what was done by Williams and Martinez [10]. They ordered their niche space based upon the mass of each species in their network. We felt that ordering based on the Trophic Level of each species followed the same line of thinking done here. Additionally, this allowed us to reuse the results made previously and add to the uniqueness of our model. Another helpful resource was a forum page on Stack Overflow [7], that outlined the process of optimizing

for multiple variables. The top post suggested making changes to all the free variables at once and then applying the Metropolis Criterion. We chose to optimize first for diet position and then for feeding range for each iteration of the simulated annealing.

Another thing that we need to discuss is the optimization of our PNM. We spent significant time making adjustments to our current algorithm, but do to the time complexity of simulated annealing and calculated likelihoods ($O(n^2)$), we were not able to get a satisfactory result. Given the chance in the future we would potentially try a whole different implementation such as something like only performing simulated annealing on a random sample of nodes rather than the entire network. Although, taking shortcuts like that could certainly decrease the end result's accuracy. Perhaps we could also use tools such as those provided through Amazon Web Services (AWS) to increase our computing capacity. Models such as these require a vast amount of computational power to be performed on large networks.

We unfortunately were not able to compare the models we created on the Messel Food Web to extant food webs. We felt that finishing and attempting to optimize the PNM were higher priorities than modern comparisons, because the PNM was our primary algorithm. We were disappointed that we did not come up with a more optimized implementation despite sacrificing extant food webs, but that was our result after hours of working on the accuracy and time complexity. This is also unfortunate because we were not able to prove the validity of our optimization of the PNM. If anything, we were able to improve upon the model results done in the study by Dunne, Labandeira, and Williams through using a unique implementation of the Probabilistic Niche Model.

V. Bibliography

References

[1] Frédéric Briand and Joel E. Cohen. "Community food webs have scale-invariant structure". In: *Nature* 307.5948 (1984), pp. 264–267. DOI: 10.1038/307264a0.

[2] Aaron Clauset. *Fundamentals of networks (Lecture 1), CSCI3352 Biological Networks*. Personal Collection of Aaron Clauset. Boulder, CO: Computer Science, University of Colorado, Boulder, 2021. URL: https://aaronclauset.github.io/courses/3352/#Schedule.

[3] Aaron Clauset. *Modular networks II: inference (Lecture 6), CSCI3352 Biological Networks*. Personal Collection of Aaron Clauset. Boulder, CO: Computer Science, University of Colorado, Boulder, 2021. URL: https://aaronclauset.github.io/courses/3352/#Schedule.

[4] Aaron Clauset. *Network representations and statistics (Lecture 2), CSCI3352 Biological Networks*. Personal Collection of Aaron Clauset. Boulder, CO: Computer Science, University of Colorado, Boulder, 2021. URL: https://aaronclauset.github.io/courses/3352/#Schedule.

[5] Jennifer A. Dunne, Conrad C. Labendeira, and Richard J. Williams. "Data from: Highly resolved early Eocene food webs show development of modern trophic structure after the end-Cretaceous extinction". In: (2015). DOI: https://doi.org/10.5061/dryad.ps0f0.

[6] Jennifer A. Dunne, Conrad C. Labendeira, and Richard J. Williams. "Highly resolved early Eocene food webs show development of modern trophic structure after the end-Cretaceous extinction". In: *Proceedings of The Royal Society* 281.1782 (2014). DOI: https://doi.org/10.1098/rspb.2013.3280.

[7] ElKamina. *Algorithm to optimize multiple variables more efficiently than trial-and-error*. URL: https://stackoverflow.com/questions/10627886/algorithm-to-optimize-multiple-variables-more-efficiently-than-trial-and-error.

[8] *Messel Pit Fossil Site*. 1995. URL: https://whc.unesco.org/en/list/720/. (accessed: 03.19.2021).

[9] Santa Fe Alliance for Science[YouTube Channel]. *Eat and Be Eaten: The Science of Food Webs - Jennifer Dunne*. 2013. URL: https://www.youtube.com/watch?v=mmS8lpILcNQ.

[10] Richard Williams and Neo Martinez. "Limits to Trophic Levels and Omnivory in Complex Food Webs: Theory and Data". In: *The American naturalist* 163 (Apr. 2004), pp. 458–68. DOI: 10.1086/381964.

[11] Richard J. Williams, Ananthi Anandanadsan, and Drew Purves. "The Probabilistic Niche Model Reveals the Niche Structure and Role of Body Size in a Complex Food Web". In: *PLoS ONE* (2010). DOI: https://doi.org/10.1371/journal.pone.0012092.

[12] Richard J. Williams and Drew W. Purves. "The probabilistic niche model reveals substantial variation in the niche structure of empirical food webs". In: *Ecology* 92.9 (2011), pp. 1849–1857. DOI: https://doi.org/10.1890/11-0200.1.