The 'Topics' of Ancient Life: Latent Dirichlet Allocation of the Fossil Record

Ryan Manzuk
Department of Geosciences
rmanzuk@princeton.edu

Abstract

Analysis of the large-scale patterns in the last ∼550 million years of animal life allows researchers to elucidate the evolutionary events that gave rise to the modern biosphere. These Analyses also contextualize the up-tick in modern animal extinctions, giving a point of comparison and offering prognostications for the future of Earth's fauna. One previous dimension reduction of the animal diversity patterns using factor analysis found that the evolution of animal life can be reduced to three sequential faunal groups. In this study, I represent the animal fossil record as a bag of words where each time bin represents a document, and each taxon represents a word making up those documents. With this framework, I conduct a Latent Dirichlet Allocation at both the family and genus levels. This analysis identifies eight major, sequential faunal groups (or topics) in the history of animal life as opposed to the three previously recognized.

1 Introduction

Much of the career of the great paleontologist Jack Sepkoski was spent compiling a database of every fossil occurrence ever published. At the completion of his project, Sepkoski had an unrivaled compendium, tracking ancient life throughout the past ∼550 million years at several taxonomic levels[5]. His database was then used for fundamental discoveries about the trends in animal diversity over time, potentially the most famous of which is the identification of five mass extinctions events in Earth's history[3]. This database was also the subject of an early application of machine learning in the Earth sciences. Sepkoski performed a factor analysis on the database and found that the diversity of animal life over time could be reduced to three covarying groups of families, which he dubbed 'evolutionary faunas[4].' As seen in figure 1, the total number of marine families at any given point in time can be mostly explained by just one of these successive faunas.

Since the turn of the century, a large collaborative effort has been made to curate Sepkoski's database, add new data, and make it available accessible to all who desire to use it. Today, the Paleobiology Database (PBDB, https://paleobiodb.org) is an online, open source platform to explore, download, and contribute new data to the compendium. This resource has been used for countless studies and is the main source of data for all research in the trends in the history of life. Although there is a great deal of researchers working on this dataset, there has not been a published study of a dimension reduction of the entirety of the animal record since Sepkoski's 1981 paper. Given the somewhat recent development of new dimension reduction techniques, such as Latent Dirichlet Allocation (LDA), there is motivation to re-examine the large-scale trends in the last 550 million years of animal life.

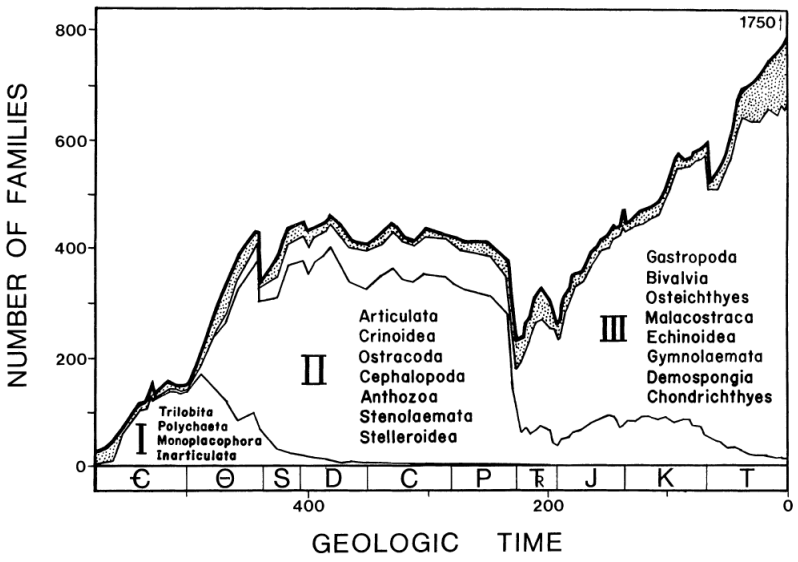

Figure 1: Resulting curve from Sepkoski's analysis, showing the covariance of three distinct evolutionary faunas over the last 550 million years. From [4]

2 Methods

2.1 Data Preparation

The dataset I used, selecting only marine taxa and excluding uncertain identifications, can be downloaded directly from the Paleobiology Database with the link included in the README file. The dataset as downloaded simply is a list of all fossil occurrences within the last 550 million years. Included with these occurrences are their taxonomic classification from phylum to genus, along with a time-binning into one of 98 geologic intervals. At the start of the analysis, I cut all fossil occurrences with incomplete taxonomic and time bin data, as well as those from the last 10,000 years because the recent is not analogous to the rest of geologic time with respect to fossil data. In order to be able to perform a Latent Dirichlet Allocation on these data, I translated them to a matrix where each time bin was given a row, and each taxon was given a column (I made two separate matrices; one at the generic taxonomic level, and one at the family level). I then filled this matrix with 1's and 0's, where 1's represent the presence of a taxon within a time interval and 0's represent the absence. Because it is impossible for a taxon to go extinct and re-evolve, whenever a taxon went through a sequence of 1-0-1—representing presence, followed by absence and reappearance—it was inferred that the taxon was actually present throughout, and the 0 was filled with a 1. For the family-level case, this matrix was 98 time bins by 4031 families, and for the genus-level case, this matrix was 98 time bins by 19,572 genera.

2.2 Latent Dirichlet Allocation

My analysis matrix, as constructed above, is analogous to a bag of words matrix where each time bin represents a document, and each taxon represents a word within those documents. It is therefore reasonable to approach this dataset with a Latent Dirichlet Allocation. This method became well developed in the early 2000's by Blei and colleagues[1], and one of its most famous applications was published by Griffiths and Steyvers in 2004[2].

In a formal description of LDA, we have four variable classes; $\beta_{1:K}$—the distribution of each topic over the vocabulary (our taxa), $\theta_{1:D}$—the topic proportions for each document (our time bins), $z_{1:D}$—the topic assignments for each document, and $w_{1:D}$—the observed words in a given document. In this model, the only observed variable is $w_{1:D}$, and all of the others are hidden or latent,

and will be solved for. The joint distribution of all these variables can then be represented as follows:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{k} p(\beta_i) \prod_{d=1}^{D} p(\theta_D) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right) \quad (1)$$

Notably, this joint distribution has two dependencies; the topic assignment of the nth word in the dth document ($z_{d,n}$) depends upon the topic proportions in that document (θ_d), and the observed nth word in the dth document ($w_{d,n}$) depends upon the topic assignment ($z_{d,n}$) and every topic ($\beta_{1:K}$).

The above dependencies are at the base of the generative process used to calculate all latent parameters. The generative process then takes the following form:

1. For each topic k = 1:K (K being a parameter defined and fed to the algorithm) draw the distribution of the topic over the vocabulary (β_k) based upon a Dirichlet distribution informed by η, the topic word prior. In the case of this study, no prior was given, so the `sklearn` algorithm defaults to a value of $\frac{1}{K}$.
2. For each document d = 1:D draw the topic proportions within the document based upon a Dirichlet distribution informed by α, the document topic prior. Again, no prior was specified, and so the default value is $\frac{1}{K}$.
3. For each word in the document n = 1:N draw its topic assignment from a Multinomial distribution of θ_d and draw the observed word from from a Multinomial distribution of $\beta_{z_{di}}$

Following the generation of these distributions, we can then calculate the posterior distribution:

$$p(z, \theta, \beta|w, \alpha, \eta) = \frac{p(z, \theta, \beta|\alpha, \eta)}{p(w|\alpha, \eta)} \quad (2)$$

Because the calculation of this exact posterior distribution is computationally intractable, the `sklearn` implementation of LDA approximates it with a simpler distribution, $q(z, \theta, \beta, |\lambda, \phi, \gamma)$ where the variational parameters λ, ϕ, γ are optimized to maximize the Evidence Lower Bound:

$$logP(w|\alpha, \eta) \geq L(w, \phi, \gamma, \lambda) \triangleq E_q[log\ p(w, z, \theta, \beta|\alpha, \eta)] - E_q[log\ q(z, \theta, \beta)] \quad (3)$$

At its core, LDA is an EM algorithm, where the distributions can be calculated generatively through the process described above based upon the dependencies elucidated in equation 1. The algorithm then iterates over this process, replacing the priors with the posteriors of previous iterations until the log likelihood of the data given the topics has been optimized. Specifically, as it is implemented in `sklearn` the user can select a maximum number of iterations of the EM step or a stopping tolerance (which encodes a lack of change when updating) has been reached. In general, I found it difficult to reach the default stopping tolerance, and that increasing the number of iterations made little difference to the final results.

3 Results

Note: The results for the family-level LDA are presented here, while the genus-level results are given in the appendix.

The first step in determining the best model for analysis through Latent Dirichlet Allocation is the selection of the number of topics that describe the dataset. I did this by training models with various numbers of topics from two to 100 and comparing the log likelihood scores of each model. In comparing these values (Fig. 3), I found that a nine-topic model best explained the data with a score of -258015.

With the proper model in hand, the most important result to visualize is how the topics vary over time. When the weight of each topic is plotted throughout each successive time bin (Fig. 3-A), we see that eight of the topics form a sequence of 10-50 million year periods where each explains

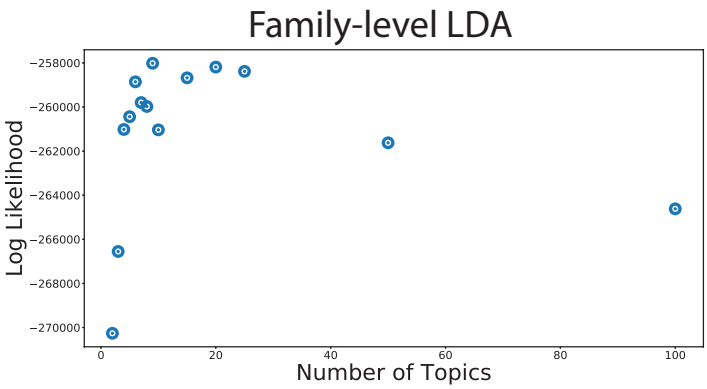

Figure 2: Log likelihood of the dataset given the topics for LDA models run at various numbers of topics on the family matrix. The peak value occurs at nine topics.

T7	T3	T1	T2	T4
Hyolithomorpha	Heteractinida	Crinoidea	Polyplacophora	Ophiuroidea
Helcionelloida	Helcionelloida	Stromatoporoidea	Crinoidea	Gymnolaemata
Lingulata	Lingulata	Paragastropoda	Rostroconchia	Actinopteri
Trilobita	Trilobita	Anthozoa	Stenolaemata	Gastropoda
Conulata	Rhynchonellata	Heteractinida	Ophiuroidea	Scaphopoda
T5	**T8**	**T6**	**T9**	
Ophiuroidea	Chondrichthyes	Bivalvia	Cephalopoda	
Asteroidea	Ophiuroidea	Reptilia	Insecta	
Osteichthyes	Anthozoa	Mamalia	Reptilia	
Maxillopoda	Gymnolaemata	Anthozoa	Chondrichthyes	
Cephalopoda	Gastropoda	Gastropoda	Bivalvia	

Table 1: The top five classes represent the heaviest-weight families of each topic. Topics are listed in temporal order, and topic 9 is low-weight throughout.

nearly all of the diversity. The transitions between these peaks represent periods of change between the faunal topics, which tend to last on the order of 50 million years. As might be expected, several of these transitions center around one of the five major mass extinction events (gray vertical lines in Fig. 3) in Earth's history.

Although there is no existing model for comparison, the topics found here make sense—in terms of both timing and make-up—when compared with knowledge of the geologic record. First of all, several faunal transitions center around mass extinctions, which are times of obvious turnover in Earth's biosphere, giving confidence that this model is picking up on actual trends. The major animal classes of the heaviest-weight families of each topic are given here in Table 3, and a more complete table is included with this submission. Like Sepkoski noted with his three faunas, the major constituents of each topic are indeed the animal groups we see dominating the fossil record during the times when their respective topics are at their peaks. Trilobites, lingulids, and stromatoporoids, which dominate museum shelves during the early periods of animal history eventually give rise to the most recent topic, which contains more familiar groups like snails, clams, and mammals.

4 Discussion and Conclusion

The topics discovered by the LDA models can be seen as akin to Sepkoski's evolutionary faunas. However, the resulting topics of an LDA allow for mixed membership, whereas the evolutionary faunas discovered through factor analysis were rigidly bound groups of animal classes. This mixed membership likely accounts for the detection of many more than just three groups in that it allows certain taxa to persist through a faunal transition, and thus does not require whole-sale turnover. One simple example of this is the transition between topics 7 and 3—the first two topics in the family-

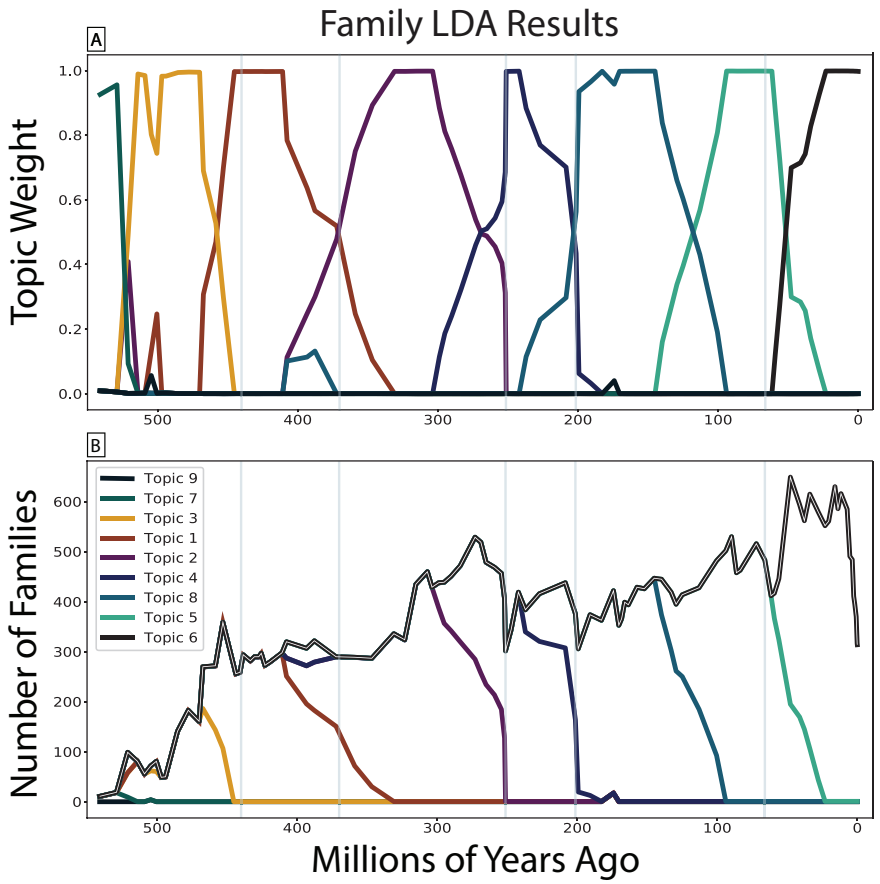

Figure 3: (A) Weighting of each topic plotted against geologic time for the family-level LDA. The weight is essentially the percent of that period's fauna that can be accounted for by each respective topic. Notably, each topic (besides Topic 9, which is low-weight at all times) covers tens of millions of years where it essentially explains the entirety of Earth's fauna. The gray vertical lines represent the five major mass extinctions, which likely cause some of the topic switches. (B) Weightings from (A) multiplied by the number of families in each time period to give the total number of families accounted for by each topic. The curves are plotted additively such that the topic 6 curve is actually the total of every topic's curve. The which line overlaying that curve thus represents the total family-level diversity through time.

level LDA. When examined at the class-level, we see that trilobites and lingulid brachiopods are both major components of each topic.

The most interesting result of this new model with more groups making up the fossil records is the recognition of faunal transitions during time periods where there are no mass extinctions. These intervals require other explanation as to the reason the animal community changed over, offering new potential focuses of research for geologist and paleontologists. For example, right at the start of the time period analyzed, the family-level LDA recognizes a short-lived, low-diversity topic that disappears approximately 520 million years ago. I would label this the Tommotian Fauna after the time period it mostly exists within, and it mostly comprises some trilobite and inarticulate brachiopod families, as well as small shelly fossils (that is the actual technical term for them). The families that make up this fauna indeed only exist for this short period of time early in animal evolution, and so this raises the question as to why there was such a faunal transition at this point. Was there an environmental change that allowed other groups of animals to take off, while the Tommotian fauna went extinct?

The next step for this project likely is to implement a dynamic or sequential topic model. The model presented here has no input of the data as a time-series (making the result of a series of sequential topics all the more compelling). Thus, if we implement a model that has prior knowledge of the the data as a time-series we may see an even more accurate result.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[3] David M Raup and J John Sepkoski. Mass extinctions in the marine fossil record. *Science*, 215(4539):1501–1503, 1982.

[4] J John Sepkoski. A factor analytic description of the phanerozoic marine fossil record. *Paleobiology*, 7(1):36–53, 1981.

[5] J John Sepkoski. A compendium of fossil marine animal families. *Contributions in biology and geology*, 83:1–156, 1992.

5 Appendix

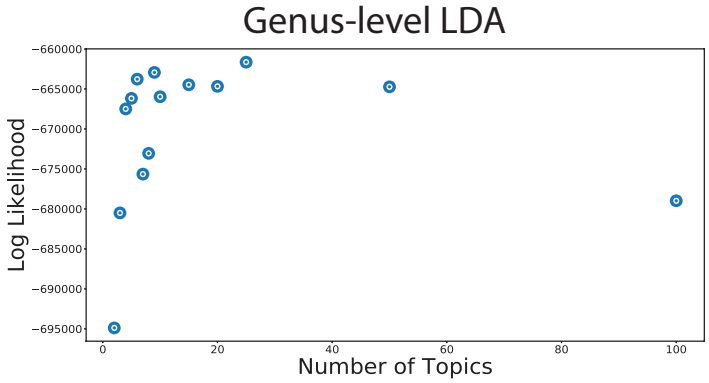

Figure 4: Log likelihood of the dataset given the topics for LDA models run at various numbers of topics on the genus matrix. The peak value occurs at 25 topics, but that model only had 9 topics of significant weight. Thus, the other peak of a nine-topic model was chosen for analysis.

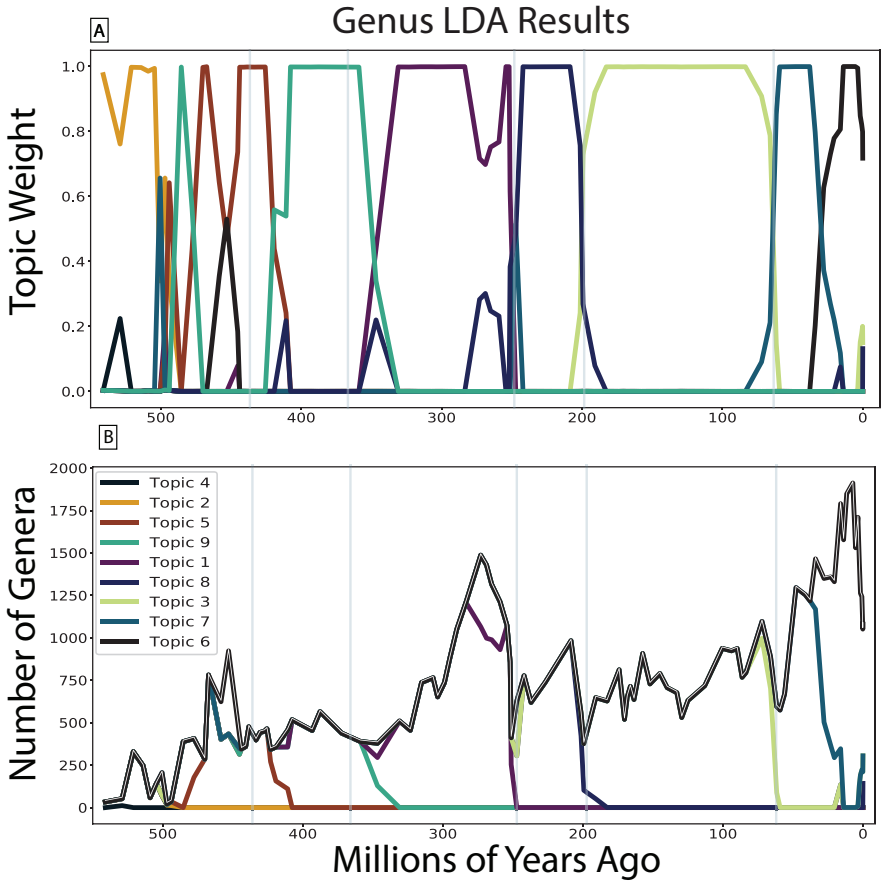

Figure 5: (A) and (B) are the same plots as Fig. 3 but for the results of the genus-level LDA. As in the family-level analysis, we see that the topics form a sequence of domination, but the time-periods are more variable in this case. Faunal transitions also occur at different points in time when compared to the family analysis.