Probability Models for Speciation and Extinction

MAS8391 Report

Naomi Hannaford

Supervisor: Dr. Tom Nye

Abstract

Species loss is an increasing problem; its consequences affect all living organisms on Earth. Therefore, modelling speciation and extinction is useful and important in contemporary research. This project explored different probabilistic models for speciation and extinction and utilised them to create methods of simulating speciation and speciation-extinction trees in R. These trees were used to investigate the contradictory results of two different papers regarding the effect of species loss on phylogenetic diversity. As I have shown, the disagreement occurred because of the use of different models of speciation. Another paper has shown that it is possible to infer speciation and extinction rates from speciation trees, even though extinction events are completely missing from such trees. This project also looked at how this surprising result was obtained. Although the reasoning behind the result is sound, it is, in practice, a difficult estimation to perform.

Contents

1 lntroduction

1.1 What is Evolution?

Evolution is the process by which inherited characteristics of organisms change over time [2]. It is responsible for the vast diversity of species in the natural world. Before we can explore evolution in more detail, we need to understand some fundamental biological terms.

When thinking about humans, we all have physical and behavioural characteristics or *phenotypes*, which make us look different and behave in different ways. Genes in our DNA, along with the environment, determine our phenotypes. For example, the eye colour of a person is determined by their DNA, which they have inherited from their parents. *Inheritance* is the process of DNA being passed from generation to generation. An example of when the environment affects a physical characteristic could be exercise changing a person's body type. A person can gain a more muscular physique by exercising a lot. Although, one can argue that a person's genes also help to determine the natural build of a person's body. The same principles apply to all living organisms, from algae to oak trees; ants to elephants; staphylococcus to salmonella. The characteristics of an individual from each species will always be determined by their DNA **and** their environment.

For a species to survive, members of the species need to reproduce successfully. The probability that an individual will reproduce can be thought of as a function of the environment and DNA, i.e.,

$$\Pr(\text{Reproduce}) = f(\text{DNA}, \text{environment}).$$

Genotypes which increase the probability of reproduction tend to be favoured over time. This idea makes sense, since organisms that are better equipped to surviving will be more likely to reproduce than the weaker individuals. For example, some strains of bacteria have become more resistant to particular antibiotics, due to specific traits determined by their DNA, which means they are more likely to survive and reproduce. It is these strains of bacteria that will start to dominate, with the ones lacking the resistant traits being destroyed by the antibiotics. Naturally, over time, the bacteria will evolve so that all strains become resistant to certain antibiotics. Bacteria reproduce better under specific conditions too, for example, optimum pH levels and temperature. This also briefly shows how probability of reproduction is affected by both environment and genetics.

DNA inheritance is not error free though. Novel genotypes can arise due to random mutations or errors made during the process of copying DNA. This is called *variation* or genetic variability and it is very important for the process of evolution, as it gives rise to new genes and therefore new physical traits.

Speciation is the process of a population of a single species splitting into two new, different species. A species is a population of organisms which cannot breed successfully

with members of another species. Note that this is a partial definition: for example, bacteria often reproduce asexually, but we still consider different species of bacteria. We now look at a hypothetical example of speciation. Imagine there is a population of black ants living in an area where there are mountains and jungle. A random mutation in one of the ant's DNA occurs, and some of its offspring are green in colour. These green ants survive better in the jungle, as they are camouflaged from their predators. However, the black ants are better adapted to surviving in the mountains. Over a long period of time, the green ants colonise the jungle area, and the black ants colonise the mountain area. Both groups of ants continue to reproduce successfully. As more variations in the genotypes of the ants occur, the two types of ants slowly become different from each other, until they are separate species, which cannot breed together. This is a very simple example of speciation.

One important type of speciation is allopatric speciation. This occurs when a population is split up due to a geographical separation, such as rising sea levels causing areas of land to be separated, or the formation of mountain ranges, due to tectonic activity. Each population may live in different types of environment, which means *selective pressures* could be different. Selective pressures are factors of the environment that mean certain traits will be favoured over others. These examples remind us of the large 'geological' timescale that evolution happens over.

1.2 Trees

1.2.1 What is a Tree?

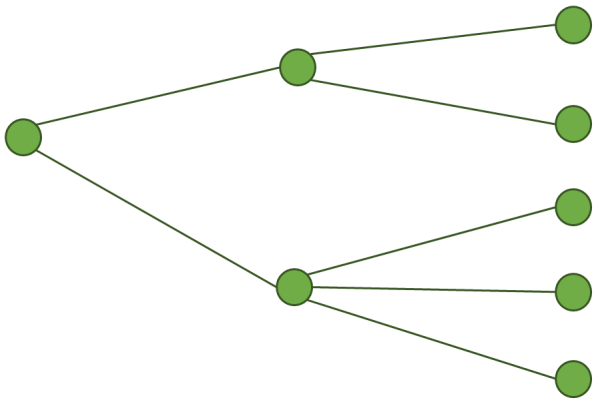

Figure 1: An example of a tree from graph theory.

We begin with some definitions from graph theory that will be useful. A graph is a set of vertices V connected by edges E. In graph theory, a forest is a graph with no cycle. A connected subgraph of a forest is known as a tree; it is a connected graph that does not have a cycle (see Figure 1 above). Trees and forests are *simple* graphs. A simple

graph does not have any *multiple edges* or *loops*. Multiple edges are edges which connect the same two vertices more than once. A loop occurs when a vertex is connected to itself. A graph with n vertices is a tree if and only if it is connected and has $n-1$ edges. We say each vertex v has a *degree* $\deg(v)$; it is the number of edges connected to it. Weighted graphs are graphs which have numbers or weights assigned to their edges. In other words, a weight can be thought of as a function, $w : E \to \mathbb{R}$.

1.2.2 Speciation Trees

The trees we shall be looking at in this report are called speciation trees or species trees. They have the properties described above in section 1.2.1. When we trace the history of different species, the branching pattern is represented by a species tree [11]. We assume all species have a common ancestor. Speciation trees have some important properties that shall now be discussed.

1. Speciation trees are **rooted**, which means they have a vertex, the root, with degree two. This root corresponds to the *most recent common ancestor (MRCA)* of all species represented by the tree.

2. Speciation trees are **binary**. This means all vertices, excluding the root, have degree one or three. If a vertex has degree one, then it is a *leaf* or *tip*. If a vertex has degree three, then it is an *internal* (or *ancestral*) vertex, i.e.,

$$\begin{aligned} \deg(v) = 1 &\iff \text{leaf} \\ \deg(v) = 3 &\iff \text{internal vertex} \end{aligned}$$

   Internal vertices represent *speciation events* or *cladogenesis* (the splitting of a population into two new clades or groups). *Terminal vertices* (or leaves) represent *extant* species. Extant means currently alive or existing.

   Although we do not consider such cases, it is worth noting that a vertex with $\deg(v) > 3$ represents a species with one ancestor and more than two descendants. This is called a *polytomy*. There are two types of polytomy; 'hard' and 'soft'. A hard polytomy is representative of a simultaneous divergence, where all the descendants evolved at the same time [11]. Uncertainty about phylogenetic relationships is indicated by a soft polytomy. In other words, the species did not all diverge at the same time, but the actual order of divergence is not fully known. Usually polytomies are considered to be soft.

3. Speciation trees are **weighted**, where weight has the same meaning as described in section 1.2.1. Their edge lengths (*branches*) represent duration between speciation events. We also define the *distance* of a leaf from the root as the sum of edge weights on edges between the root and leaf.

4. Speciation trees are **ultrametric/clocklike**. The leaves of a speciation tree are all equidistant from the root, which means they are *ultrametric* trees or *dendrograms*. The property of ultrametricity means we have a well-defined notion of time from the root to any point on tree [11].

5. Speciation trees are often **leaf labelled**. Usually their leaves are labelled with the species' names.

A shorthand for representing speciation trees is called *Newick strings*, which use nested parentheses. Each internal vertex of a species tree is represented by a pair of parentheses containing all the descendants of that particular vertex. Newick strings enable us to describe a tree without having to draw it. For example, Figure 2 below can be represented by the Newick string (((2,3),4),1). The edge lengths (time between speciation events) can also be included into Newick string notation by having a colon after the species' name and then the edge length, i.e., species' name: edge length. In section 2.3.2, we shall demonstrate how Newick strings can be used in our simulations of speciation and extinction to create species trees, in R. For more information on speciation trees and their properties see [11].

Figure 2 below shows how speciation trees will be drawn in this report. The large dot marked on the tree, represents the point where the root/MRCA first splits into two new species. There is a time axis on the bottom, which starts at the root and ends at present time. This tree has four extant species and is leaf labelled, with labels 1, 2, 3 and 4. Note that we are free to 'rotate' vertices, in other words, the same tree can be represented in several ways. In Figure 2 we can 'rotate' vertex 4 and vertices $(2, 3)$ or swap vertex 2 and vertex 3, and obtain the trees in Figure 3. These trees are equivalent to the tree shown in Figure 2.

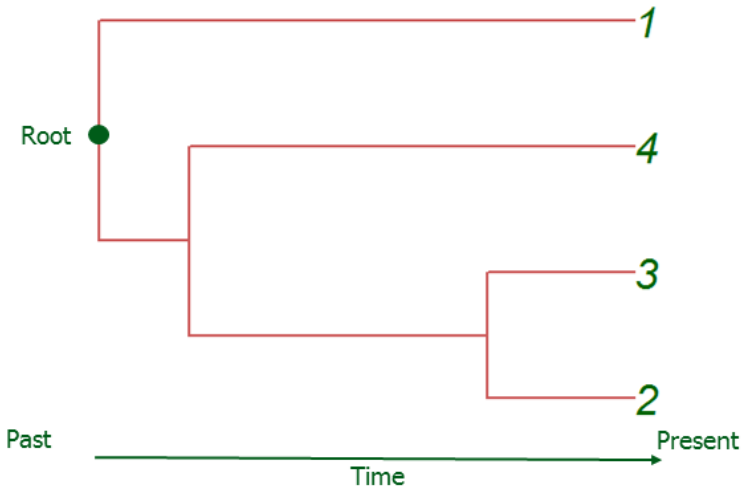

Figure 2: An example of a tree with four extant species.

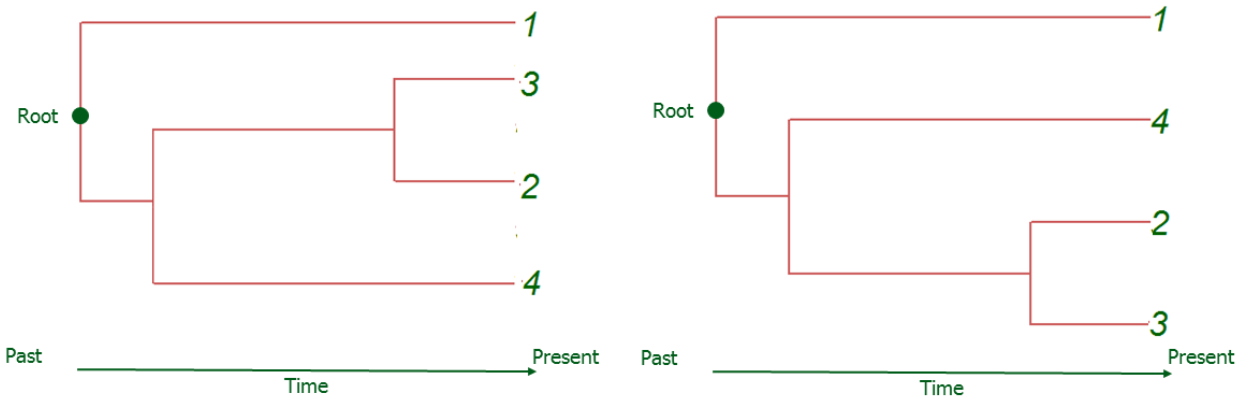

Figure 3: Equivalent speciation trees to Figure 2 after two different rotations of vertices, swapping $(2, 3)$ with 4 (left) and swapping 2 and 3 (right).

1.2.3 Speciation-Extinction Trees

Speciation-extinction trees are very similar to speciation trees and have all the properties described in section 1.2.2 above, however, they also have leaves at some points in the past, between the MRCA and the currently existing species. These leaves represent extinct species or extinction events. Below is an example of a speciation-extinction tree.

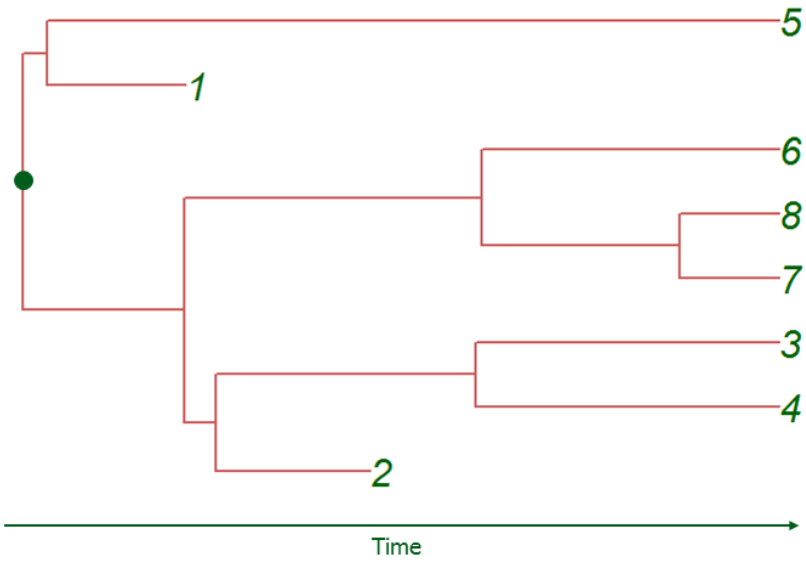

Figure 4: A speciation-extinction tree with 6 extant species, 7 ancestral species (including the MRCA) and 2 extinct species (1 & 2).

1.2.4 Tree Topology and Shape

If we keep the leaf labels of a speciation tree and discard the branch lengths from a species tree then we obtain the *tree topology*. For four species there are 15 topologies in total. The number of possible topologies for a rooted species tree with n extant species is

$$\begin{aligned} T_n &= (2n-3)(2n-5)...(3)(1) \\ &= (2n-3)!!. \end{aligned}$$

To understand why this is true, we consider a counting argument, where we build up all of the possible trees by adding species one by one, in a predetermined order. Say we have all the possible trees with n species and we then add species $n+1$ to each of them, in all possible positions. By doing this, we will create all possible trees with $n+1$ species, without repetition. The new species cannot be connected to an existing interior vertex, since speciation trees are binary, as described in section 1.2.2. It must be connected to a new vertex and this is placed in the middle of an existing edge.

To see why this produces all possible species trees, we consider two operations, which are the inverse of each other:

1. Add species k to a $k-1$ species tree.
2. Remove species k from a species tree containing $k-1$ species.

Suppose we have a particular n species tree and we remove successively species n, $n-1$, ... $k+1$. Once species $k+1$ has been removed, we are left with a species tree with k species. As the removal operation is the inverse of the addition of species, there must exist a certain sequence of positions to add species $k+1$, $k+2$,...n onto that k species tree to yield the original n species tree. Moreover, no other tree with k species can obtain that particular n species tree (by adding on the $n-k$ missing species). Another k species tree cannot yield the original n species trees because that tree would also be reached by removal of the $n-k$ species and this is not possible, as the same sequence of removals cannot result in two distinct trees. Hence, any n species tree can be obtained from one, and only one, k species tree.

Thus, each possible sequence of addition to a species tree with k species leads to a different n species tree, and all possible trees can be created in this way. The number of ways in which we can add a species to a tree is the same as the number of edges. There are three edges in a tree with two species. Addition of a new species results in adding a new vertex and two new edges. After adding the third species to one of the three possible positions, there are then five possible places for the the fourth species to be added to, seven places for the fifth and so on. There are $2n-3$ places that the species n can be added to. Hence, we have shown that there are $3 \times 5 \times 7 \times ... \times (2n-3) = (2n-3)!!$ possible, distinct ways to generate a tree with n species [5], as stated above.

The *shape* of a species tree is what remains if we ignore edge weights and leaf labels. When drawing the shapes of a species tree, we follow the convention of 'leaves to the top', i.e., if a branch does not split (if a speciation event does not occur) then that leaf moves to the top, whereas a speciation event is kept at the bottom of the tree. This is because some species tree shapes can be the same via rotation of ancestral vertices (see Figure 5 below). Following this rule enables us to identify equivalent shapes. For four species, there are two shapes the tree can take.

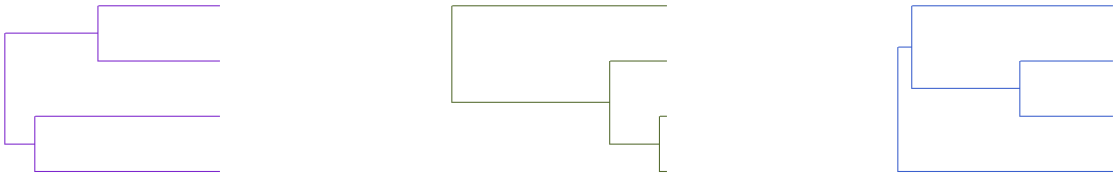

Figure 5: Tree shapes for four species. The blue tree does not follow the rule of leaves to the top and is the same shape as the green tree.

1.2.5 Representing Species Trees as Data Frames in R

I have developed an R data structure for representing species trees and we shall utilise this when carrying out simulations later in section 2. The data frame has seven columns (`time`, `time from parent`, `parent`, `extant`, `child 1` , `child 2` and `extinct`) and each row represents a vertex in a tree. `Time` is the furthest point in time from the root (MRCA) when the species existed. The `time from parent` column records the time since the ancestral speciation event. The `parent`, `child1` and `child 2` columns keep track of which species (the `parent`) split to create two new species (`child 1` and `child 2`) and in these 'child' columns row numbers are recorded. These columns keep track of the ancestry for each species. If a species is currently alive, then its `extant` column entry will be `TRUE`. When a species becomes extinct then its `extinct` entry will be `TRUE` and its `extant` entry will be `FALSE`. An internal/ancestral vertex is represented by the presence of `FALSE` in both the `extant` and `extinct` columns. Below, in Output 1, is an example of an R data frame representing a species tree (Figure 6).

	time	timeFromParent	parent	extant	child1	child2	extinct
1	0.4697740	0.00000000	NA	FALSE	2	3	FALSE
2	0.5109237	0.04114967	1	FALSE	4	5	FALSE
3	0.5485330	0.07875894	1	FALSE	6	7	FALSE
4	0.6343637	0.12344006	2	FALSE	8	9	FALSE
5	0.6908246	0.17990092	2	TRUE	NA	NA	FALSE
6	0.5596362	0.01110321	3	FALSE	NA	NA	TRUE
7	0.6908246	0.14229164	3	TRUE	NA	NA	FALSE
8	0.6908246	0.05646086	4	TRUE	NA	NA	FALSE
9	0.6579336	0.02356989	4	FALSE	10	11	FALSE
10	0.6908246	0.03289097	9	TRUE	NA	NA	FALSE
11	0.6908246	0.03289097	9	TRUE	NA	NA	FALSE

Output 1: Example of a data frame representing a species tree with five extant species and one extinction event (row 6).

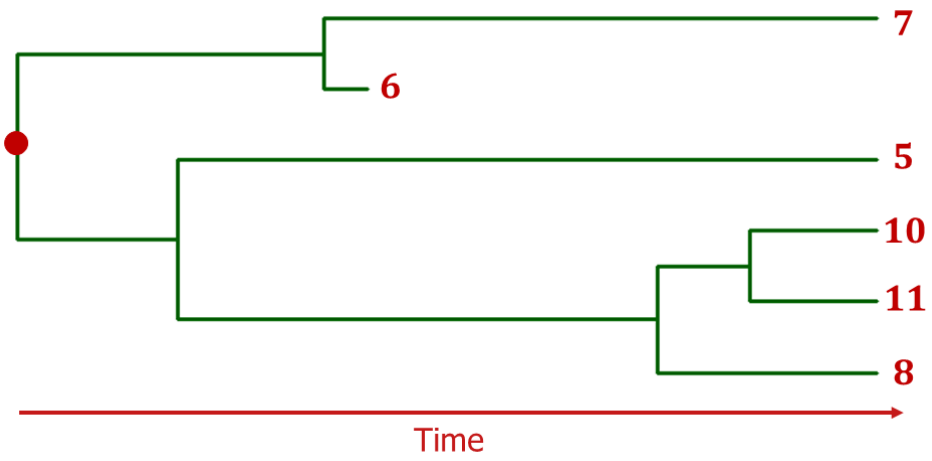

Figure 6: Species tree represented by the data frame in Output 1 above.

1.3 Aims of this Report

In this report, we wish to explore the distribution of species trees, by looking at different models currently available, and simulating species trees in R using these models. We shall also be looking at some interesting and controversial findings from a paper written by Nee et al. concerning species loss and conservation [9]. By using simulations we shall try to understand these results better. After analysing the details of this paper, we shall then go on to look at the findings of another paper by Steel et al. which argues against Nee's findings [13]. Finally, we shall look at a result from a paper by Harvey et al. [6]. The result says it is possible to estimate both speciation and extinction rates by using only speciation trees.

2 Modelling Evolution

Before we look at models for species trees, we need to gain some understanding as to where these models originate from. Therefore, we shall look at models for population sizes before describing how to use these to define models for speciation trees.

2.1 The General Birth-Death Model

Birth-death models are used to represent population dynamics. They are a collection of random variables $X(t)$ taking values in $S = \{0, 1, 2, 3, ...\}$, as population size is discrete, and where $t \in \mathbb{R}$. Sometimes there is an upper limit on the population size and then $S = \{0, 1, 2, ...N\}$. Birth-death processes are *Markov processes*. A Markov process is a stochastic model with the Markov property [1]. If a process has the Markov property, then it means that future states only depend on the current state, i.e.,

$$\Pr(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_0 = x_0) = \Pr(X_n = x_n | X_{n-1} = x_{n-1}). \tag{1}$$

In the context of birth-death models, it means that a change in population size is only dependent on the current population size. In a small interval of time $(t, t + h]$, the population changes by $+1$ ('birth'), by -1 ('death') or remains the same.

Let $X(t)$ be the population size at time t with $X(0) = n$, where n can take any positive integer value. The general birth-death model has birth and death rates proportional to population size, so that when $X(t) = k$, the birth rate is $\lambda \times k$ and the death rate is $\mu \times k$, with λ, μ constant. The interevent time has an exponential distribution, $\text{Exp}((\lambda + \mu)k)$. The probability of a death at any time is $\frac{\text{death rate}}{\text{total rate}} = \frac{\mu}{\lambda+\mu}$. The probability of a birth occurring is $\frac{\text{birth rate}}{\text{total rate}} = \frac{\lambda}{\lambda+\mu}$.

2.2 The Simple Birth Model

The simple birth model is used to represent population size when there are births but no deaths. There is a constant rate of birth.

Let $X(t)$ be the population size at time t with $X(0) = n$, where n can take any positive integer value. When $X(t) = k$, the birth rate is $\lambda \times k$, where λ is constant. The interevent time, the time between each birth has an exponential distribution, $\text{Exp}(\lambda k)$, when $X(t) = k$.

For a simple birth process $X(t)$, with initial condition $X(0) = n$,

$$\begin{aligned} \text{E}[X(t)] &= ne^{\lambda t} \\ \text{Var}(X(t)) &= n(e^{2\lambda t} - e^{\lambda t}). \end{aligned}$$

More information on the general birth-death model and the simple birth model (including the reasoning for the above result) can be found in [1].

2.3 Speciation-Extinction Tree Models Based on the General Birth-Death Model

Now we transform the general birth-death model into a model for speciation-extinction trees. With no extinction ($\mu = 0$), this is called the Yule Model. Let $L(t)$ be the set of species extant at time t and $X(t) = |L(t)|$, with $X(0) = 1$, as we shall assume we are starting with one species, the MRCA of all species. When a birth occurs, a species is chosen uniformly at random from $L(t)$ to split, so $|L(t)|$ increases by 1. However, when a death occurs, a species is chosen uniformly at random from $L(t)$ and removed, and $|L(t)|$ decreases by 1. The probabilities of a birth or a death occurring are the same as described in section 2.1. When $X(t) = k$, we have the same interevent time described in section 2.1. These interevent times determine the edge lengths of the species trees. At some point in the past we have the MRCA and we move forward in time to the present. We fix the maximum number of leaves to be n, an upper limit on the number of extant species, which determines when the birth-death process stops. In section 2.3.2, we look at implementing the model in R and plotting simulated speciation-extinction trees. It follows that the simple birth model can be used as a basis for a model for speciation trees, since it is just a type of birth-death model.

2.3.1 The General Birth-Death Model Simulation in R

In R, there are three steps for generating species trees. First, we use a function based on the general-birth death model to create a data frame with the structure outlined in section 1.2.5. The second step is to convert this data frame into a Newick string. Once we have the data in the form of a Newick string, we can then plot the species tree. The species tree will always be binary because the model only splits single species into two.

The following algorithm is used to create the data frame. Our function has the parameters N (number of terminal vertices we would like), λ (speciation rate) and μ (extinction rate).

1. Start with one entry (root or MRCA) in the data frame with:

   `time` $= 0.0$, `time from parent` $=$ `0.0`, `parent` $=$ `NA` (since it is the root), `extant` $=$ `TRUE`, `child 1` $=$ `NA`, `child 2` $=$ `NA` and `extinct` $=$ `FALSE`.

2. Set up a while loop. Steps $3-7$ are repeated until there are 0 or N extant species in our data frame.

3. Generate an interevent time $dt \sim \text{Exp}((\lambda + \mu)k)$, where k is the number of extant species.

4. Generate a random uniform variable u.

5. Randomly sample from the extant species (i.e., rows with `extant == TRUE`).
   (a) If $u \leq \frac{\mu}{\lambda+\mu}$, then we 'kill off' the randomly sampled species (set `extant` = `FALSE` and `extinct` = `TRUE`). Set $k = k - 1$.
   (b) If $u > \frac{\mu}{\lambda+\mu}$, a speciation event occurs. For the sampled species, we set `extant` = `FALSE` and `child 1` to the number of rows currently in the data frame plus one and `child 2` to the number of rows plus two. Two new rows are also created with `time` = `time`, `time from parent` = 0.0, `parent` = row number of sampled species, `extant` = `TRUE`, `child 1` = `NA`, `child 2` = `NA` and `extinct` = `FALSE`. Set $k = k + 1$.

6. Add dt onto all the extant species' times.

7. Add dt to the `time from parent` entries of the extant species.

8. Finally, label extant species uniformly at random, without replacement, from $\{1, ..., N\}$.

The algorithm for the simple birth model is very similar to the algorithm above. However, step 4 is not executed and step 5(b) is executed every time. The `extinct` column can also be omitted from the data frame.

2.3.2 Converting to a Newick String and Plotting the Species Tree

To convert our data frame into a Newick string, we write a function which checks and assigns a string to each row in our data frame, starting from the bottom. If the row (representing a vertex) has no descendants, then it is a terminal vertex and its string is recorded as:

`string = species' name:time from parent`, for example, 1:0.25.

However, if the vertex does have descendants then its string is:

`string = (child1string,child2string):time from parent`, for example,

(4:0.25,5:0.25):0.1.

Since we work up from the bottom, the string associated to the first row (MRCA) in the table is then the Newick string for the entire tree. We can use functions within the library `ape` to produce a tree plot, using this string. `ape` is a package in R, which is used to analyse phylogenetics and evolution; its functions include the writing, plotting and manipulation of phylogenetic trees and many more. Figure 4, in section 1.2.3, was plotted using the steps described here with the parameters $N = 6, \lambda = 2$ and $\mu = 1$.

2.3.3 The Yule Model and Probability Distribution of Trees

If we consider the birth model (Yule model) then it might naively be expected that the distribution on tree shape under the model is uniform. This is not the case. As discussed in section 1.2.4, there are two different tree shapes for a tree (see Figure 5) with four extant species. If we label the terminal vertices in all the possible, distinct ways to obtain the tree topologies, we find there are three possible topologies for the purple tree in Figure 5. The probability of a purple tree occurring is $1/3$, so therefore, the probability of each of its three topologies occurring is $1/9 \neq 1/15$ (there are 15 possible topologies for a speciation tree with four extant species). Hence, the distribution of the tree topologies is not uniform.

2.4 The Coalescent Model

2.4.1 What is the Coalescent Model?

The coalescent model is used for modelling random trees. However, technically it does not represent the process of speciation. Coalescent trees are used as models in population genetics, but we shall explore their use in modelling speciation anyway, as they are commonly used as speciation models. In particular, we want to compare the coalescent model with the simple birth model and, in section 2.4.2, we shall prove that the two models produce identical distributions for tree shapes and topologies. For the coalescent model, we start with k species and choose two of these species uniformly at random to coalesce (join together) to give $k-1$ species. We continue to choose two species to coalesce uniformly at random from the remaining number of species until there is just one left. This model differs from the simple birth model, as it works backwards in time until the MRCA is reached, rather than starting from the MRCA and moving forwards in time. The interevent times follow an exponential Exp $\binom{k+1}{2}$ distribution and a brief explanation for this is giving in section 2.4.4.

2.4.2 The Wright-Fisher Model of Genetic Inheritance

To understand the coalescent model more, we shall consider the Wright-Fisher model for genealogical relationships. This model is used to track the inheritance of genes one from generation to the next [7]. There are two types of models, the *haploid* reproduction model and the *diploid* reproduction model. Haploid refers to a single set of unpaired chromosomes or genes, whereas diploid refers to a set where there are two complete pairs of chromosomes or genes. Both models have a constant population size of $2N$ genes either corresponding to $2N$ haploid or N diploid individuals. For the haploid model, each gene in generation $k+1$ is a copy of a gene from generation k. All $2N$ genes in generation $k+1$ are sampled with equal probability, with replacement, from the previous generation. Since the sampling is with replacement, it is possible for a gene in generation k to not

have any descendants, meaning its *lineage* dies out at that generation. Lineage is a term for a group of individuals (or species) which have descended from one ancestor. In this case, individuals are genes. A gene can also have many descendants too. The genes in generation k can be sampled from generation $k-1$ and so on. This concept of moving backwards in time is synonymous with the coalescent model. We shall not consider the diploid method in great detail. It is similar to the haploid model, except pairs of genes are sampled, one from a female population and one from a male population.

Diagrams of the Wright-Fisher model can be drawn for a constant population size and fixed number of generations. Genes can then be traced back to a MRCA in the earliest generation. The patterns produced by doing this are trees, as demonstrated in Figure 7 below. These gene trees are often used to represent species trees and the coalescent model detemines a distribution on a set of species with N leaves. It is worth observing that the gene trees which directly come from the Wright-Fisher model have *multifurcations*, i.e., one vertex in such a tree could split into more than two vertices. The coalescent model for species trees only involves bifurcations, as required (see section 1.2.2).

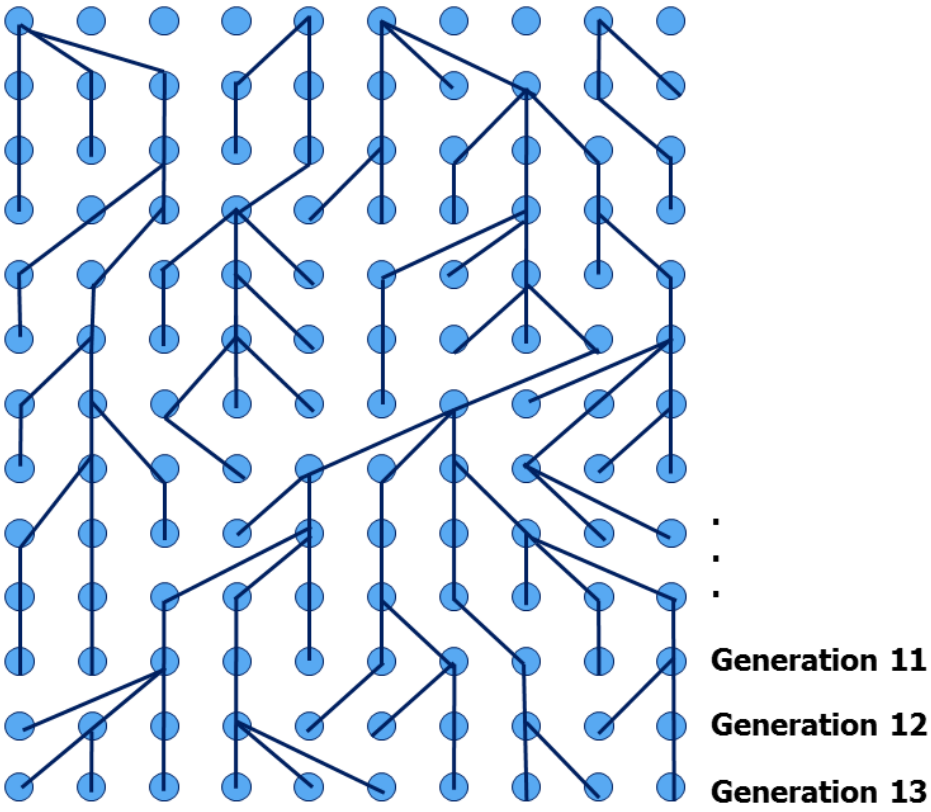

Figure 7: The haploid Wright–Fisher model with ten genes applied for thirteen generations. Starting from the top row, the model is applied twelve times.

The Markov property (explained in section 2.1) is important in both population genet-

ics and coalescent theory. In genetics, it is logical to assume that the probability of something happening, for example, finding a common ancestor, depends on only the current state of the process. For discrete time, the Markov property is associated with the geometric distribution. When we measure time continuously, we use the exponential distribution as an approximation, i.e., the waiting time until two genes share a common ancestor is exponentially distributed. See [7] for further information on the Wright-Fisher model and coalescent model.

2.4.3 Equivalence to the Simple Birth Model

It can be shown that the coalescent model distribution on species tree shape is actually the same as the simple birth (Yule) model distribution. Let P_n be the simple birth model distribution and let Q_n be the coalescent model distribution. Our claim is that,

$$P_n(T) = Q_n(T), \text{ for all tree shapes } T \text{ on } n \text{ leaves}, \tag{2}$$

where $P_n(T) = \Pr(\text{Simple birth tree has shape } T)$ and $Q_n(T) = \Pr(\text{Coalescent tree has shape } T)$.

We shall prove this by induction. For the case $n = 2$, we know that $P_n(2) = Q_n(2)$, as there is only one tree shape possible when there are just two leaves, see Figure 8 below. Now we assume that $P_n = Q_n$ is true and consider $n + 1$.

Figure 8: The only possible tree shape with $n = 2$.

Fix a tree shape T on n leaves (unlabelled). The probability that this tree occurs in the simple birth model is the same as it occurring in the coalescent model, that is $P_n(T) = Q_n(T)$, because $P_n = Q_n$. Then we consider $n + 1$ for both models.

Coalescent on $n + 1$ species/leaves: In this case, we consider leaves as unconnected vertices, as in not connected to all other vertices. We join two of the $n+1$ leaves and let these two connected leaves now represent one leaf, so now there are n 'leaves'. Suppose the tree on these remaining n vertices is T (if we were to continue coalescing to get tree T). We now have T' with $n + 1$ leaves.

Simple birth on $n + 1$ species/leaves: T' is obtained by joining the tree shown in Figure 8 uniformly at random to one of the tips of T, which corresponds to splitting one of the tips of T.

Since $P_n(T) = Q_n(T) \Rightarrow P_{n+1}(T') = Q_{n+1}(T')$. This holds for any T'.

□

2.4.4 The Algorithm for The Coalescent Model

The Wright-Fisher model measures time in discrete units. As discussed in section 2.4.2, the interevent times in the Wright-Fisher model have an approximate geometric distribution. The continuous interevent time approximation is given by an Exponential distribution. That is,

$$dt \sim \text{Exp}\left(\binom{k}{2}\right).$$

As we have shown in section 2.4.3 that the simple birth model and coalescent model give the same distribution for tree shape, we can use an algorithm similar to the one in section 2.3.1, with the interevent time distributed as described above. However, $\binom{k}{2} = \frac{k(k-1)}{2}$. Hence, the interevent time for the algorithm can be expressed as

$$dt \sim \text{Exp}\left(\frac{1}{2}k(k-1)\right). \tag{3}$$

The algorithm for simulating speciation trees, using the coalescent model, is as follows:

1. Start with one entry (root or MRCA) in the data frame with:

   `time` = 0.0, `time from parent` = 0.0, `parent` = `NA` (since it is the MRCA), `extant` = `TRUE`, `child 1` = `NA` and `child 2` = `NA`.

2. Set up a while loop. Steps $3 - 6$ are repeated until there are N extant species in our data frame.

3. Calculate $\rho = \frac{1}{2}k(k-1)$, where k is the number of extant species.
   (a) If $\rho > 0.00001$, $dt \sim \text{Exp}(1, \rho)$.
   (b) Otherwise, $dt = 0.0$. This step overcomes the problem that arises for $k = 1$. There does not exist an Exp(0) distribution.

4. Randomly sample from the extant species and set `extant = FALSE`. Set `child 1` to total number of species plus one and set `child 2` to the total number of species plus two. Two new rows are also created with `time` = `time`, `time from parent` = 0.0, `parent` = row number of sampled species, `extant` = `TRUE`, `child 1` = `NA`, `child 2` = `NA` and `lost = FALSE`. Set $k = k + 1$.

5. Add dt onto all the extant species' times.

6. Add dt to the `time from parent` entries of the extant species.

7. Label extant species uniformly at random, without replacement, from $\{1, ..., N\}$.

2.5 Comparison of the Simple Birth Trees and Coalescent Trees

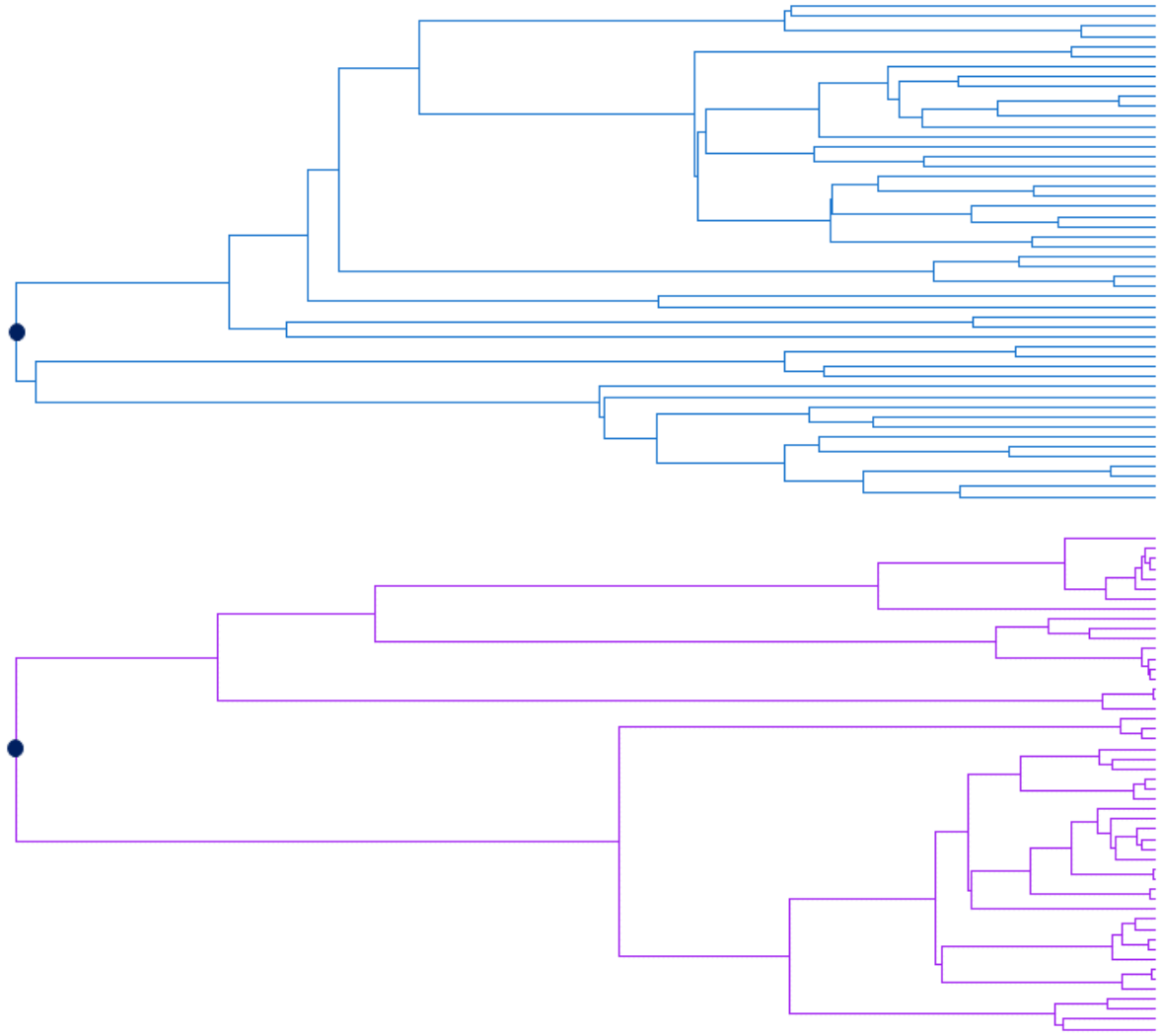

Figure 9: Birth tree (blue) and coalescent tree (purple) with 50 extant species.

Although we have shown that the two models for the distribution on tree shapes are equivalent (section 2.4.3), it is worth noting that the trees produced using the two models differ when we consider the branch lengths. Figure 9 above demonstrates the differences between the two models. The simple birth tree has longer interevent times (edges) after the MRCA, which shorten after each speciation event. Meanwhile, the coalescent tree tends to have lots of short interevent times, usually close to the leaves. This difference occurs because the interevent times are distributed slightly differently, as shown in section 2.4.1. It will be important to remember this as we look at the findings of some papers in section 3.

3 Species Loss and Phylogenetic Diversity

3.1 Controversy in Research

About 1.75 million species have been identified on Earth so far, although scientists estimate that 13 million actually exist. Unfortunately, species loss is an increasing problem. Recently, it has been reported by the International Union for Conservation of Nature (IUCN) that over 20% of vertebrates have become extinct and '16% to 33% of the remaining vertebrate species are categorised as globally threatened' [15]. It is believed that one of the main causes of species loss is the global appetite for resources. In other words, destructive human activity is the cause. Overexploitation, such as intensive farming, deforestation and mining, can lead to habitat loss. This combined with invasive species, and 'anthropogenically driven climate change' [10] is leading to devastating results. Species loss is a threat to *biodiversity*. Biodiversity is the variety of life on Earth and it is something which all species need to survive. Since species loss has been a concerning issue for some time, a lot of research has been carried out in the area.

In this section of the report, we shall be exploring the findings of two papers. The first paper, by Nee et al., claims that **'...approximately** 80% **of the underlying tree of life can survive even when approximately** 95% **of species are lost...'** [9]. However, a paper by Steel and Lambert disagrees with this statement completely; they say that **'...the loss of** 95% **would lead to the loss of more than** 84%**'** of the diversity [13]. Clearly there is a strong disagreement between the two papers. Therefore, the aim of this section is to look at how these two contradicting results were found and see why this is the case, by carrying out simulations in R and using the models described in section 2.

3.2 Phylogenetic Diversity

3.2.1 What is Phylogenetic Diversity?

Now we are going to investigate how losing extant species in a speciation tree affects the *phylogenetic diversity* (PD), by replicating the research carried out in the two papers. In general terms, phylogenetic diversity is a measure of biodiversity, which takes into account phylogenetic difference between species. With respect to speciation trees, PD is the sum of all the branch lengths in the tree, i.e., it is the total amount of evolutionary history represented by the tree [10]. If we lose an extant species in a speciation tree, then we lose its branch length, as we cannot estimate an interevent time for a species whose existence we are unaware of.

Edge lengths on tree are usually not on an absolute scale. Therefore, we typically report proportion of edge length loss when extant species are removed. A loss of a species with

a longer corresponding branch length will result in a greater loss of phylogenetic diversity than if a species with a shorter branch length is lost. If a parent species' children are both lost, then the branch length of the parent species' is also lost, as the only evidence of its existence would be the presence of at least one of its children.

Figure 10 below illustrates an example of PD loss in a speciation tree. We can see that the loss of both species E and F results in the loss of their corresponding edge lengths and their parent edge length. The loss of species A and C results in the loss of only their corresponding branch lengths.

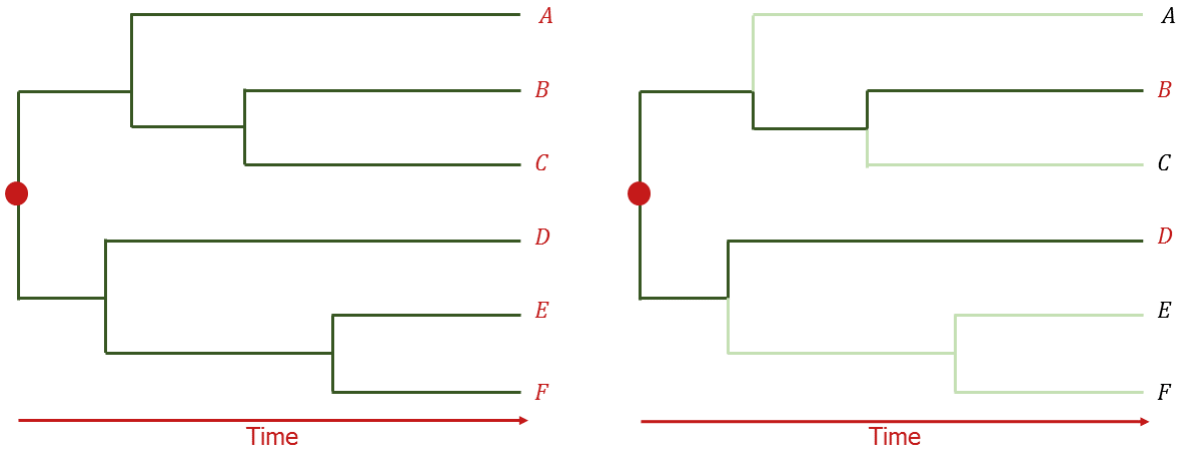

Figure 10: A speciation tree with six extant species before and after the loss of four species; A, C, E and F (in black text).

Measuring loss of phylogenetic diversity is not simple [13]. There is a complicated relationship between tree shape and edge lengths. Consider Figure 11 below. For the left hand tree, the loss of species X and Y results in a larger loss of phylogenetic diversity than if we were to lose species A and B. So, although losing the shape in Figure 8 (section 2.4.3) also means losing the parent edge length, it does not necessarily mean that PD loss is greater than the loss achieved when losing two edge lengths not forming this shape are lost. For the right hand tree, if we compare the loss of species A and B with the loss of species X and Y, we see that losing species A and B would lead to a larger loss in phylogenetic diversity. This shows us that the length of the parent edge is also important too. This example helps to explain why estimation of PD loss is a complex problem to solve.

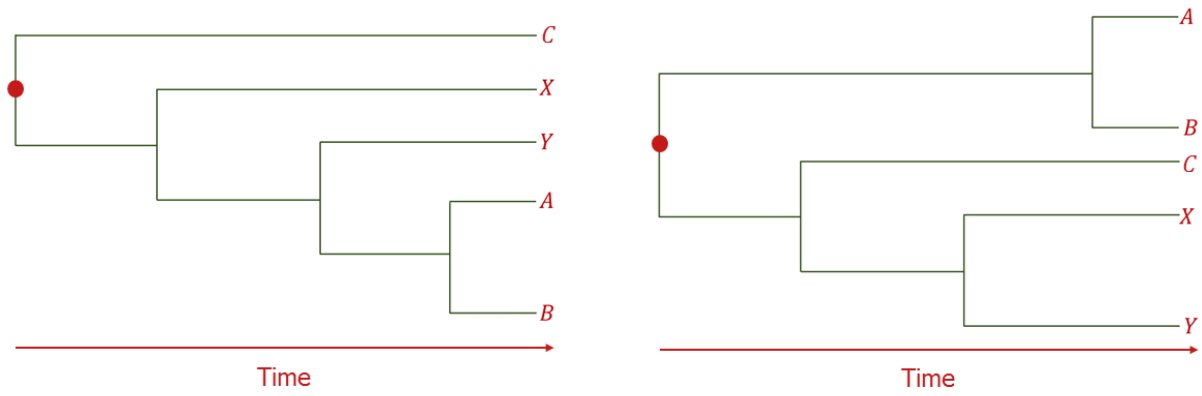

Figure 11: Two speciation trees with 5 extant species, helping to demonstrate the complexities of measuring loss of phylogentic diversity.

3.2.2 Measuring Loss of Phylogenetic Diversity Using Simulations

We begin by simulating a speciation tree in R, using either model defined in section 2.4. Next, we create a function that deletes k of the extant species, chosen uniformly at random, without replacement. The function also deletes all its corresponding branch lengths and any other internal edges that are lost due to loss of both descendant species. Then we write a function that calculates the proportion of the phylogenetic diversity saved when k species are deleted.

3.3 Papers' Findings

3.3.1 Field of Bullets Model

The *Field of Bullets* model is a model for species loss. It is the idea that all species have the same probability of becoming extinct at any time and that extinction is random (and thus, due to stochastic effects) [9,12,13]. The model does not take into consideration an organism's adaptability nor capability of surviving. The name 'Field of Bullets' comes from an analogy that all species are out in a field and 'bullets' can hit them at random.

Both Nee [9] and Steel [13] use the Field of Bullets model in their papers, however, they define and use the model in different ways. In Steel's paper, the Field of Bullets model is used in such a way that each species (represented by terminal vertices) has the probability p of being saved, the *sampling probability*, so they have the probability $1-p$ of being killed. Then every leaf (terminal vertex) is independently removed with probability $1-p$ and p is set to $\frac{k}{n}$, where n is the number of terminal vertices and $0 < k < n$. Therefore, the expected number saved will be k, but this may not be the

actual number saved in a simulation. Nee uses the Field of Bullets model such that **exactly** k out of the n species are saved.

3.3.2 Equation for PD Preserved

In Nee et al.'s paper, a formula is derived which approximates the PD preserved, denoted PD_P, dependent on how many extant species are saved in a speciation tree. This formula is based on the Field of Bullets model (section 3.3.1) and is only true for speciation trees established from the coalescent model, as we shall show in some simulation results in section 3.4.3. The approximation is given by

$$PD_P \approx \frac{\log(k-1)+C}{\log(n-1)+C}, \tag{4}$$

where k is the number of species saved and $k > 1$, n is the total number of species in the tree and C is Euler's constant (≈ 0.577) [9].

The derivation of formula (4) begins with considering the interevent time between vertices i and $i+1$. Since the interevent time is exponentially distributed, we know that the mean interevent time is proportional to $\frac{1}{i(i+1)}$. There are $i+1$ lineages between vertices i and $i+1$. Thus, this interval contributes $\frac{i+1}{i(i+1)}$ to the total phylogenetic diversity. Therefore, the total phylogenetic diversity of a speciation tree, where k species has been saved is

$$\sum_{i=1}^{k-1} \frac{i+1}{i(i+1)} = \sum_{i=1}^{k-1} \frac{1}{i} \approx \log(k-1)+C,$$

where C is Euler's constant. Similarly, the phylogenetic diversity for the original speciation tree, where no species have been lost, is

$$\sum_{i=1}^{n-1} \frac{i+1}{i(i+1)} = \sum_{i=1}^{n-1} \frac{1}{i} \approx \log(n-1)+C,$$

where C is Euler's constant. Hence, the proportion of phylogenetic diversity preserved when k species have been saved is approximated by formula (4).

Nee et al. tested their formula by carrying out simulations of coalescent trees, for various values of k and n, but did not explicitly report their simulation results. We shall look at similar simulations in section 3.4.3 to see how accurate their findings were. We are also interested in PD lost and the approximation for PD lost is

$$PD_L \approx 1 - PD_P \tag{5}$$

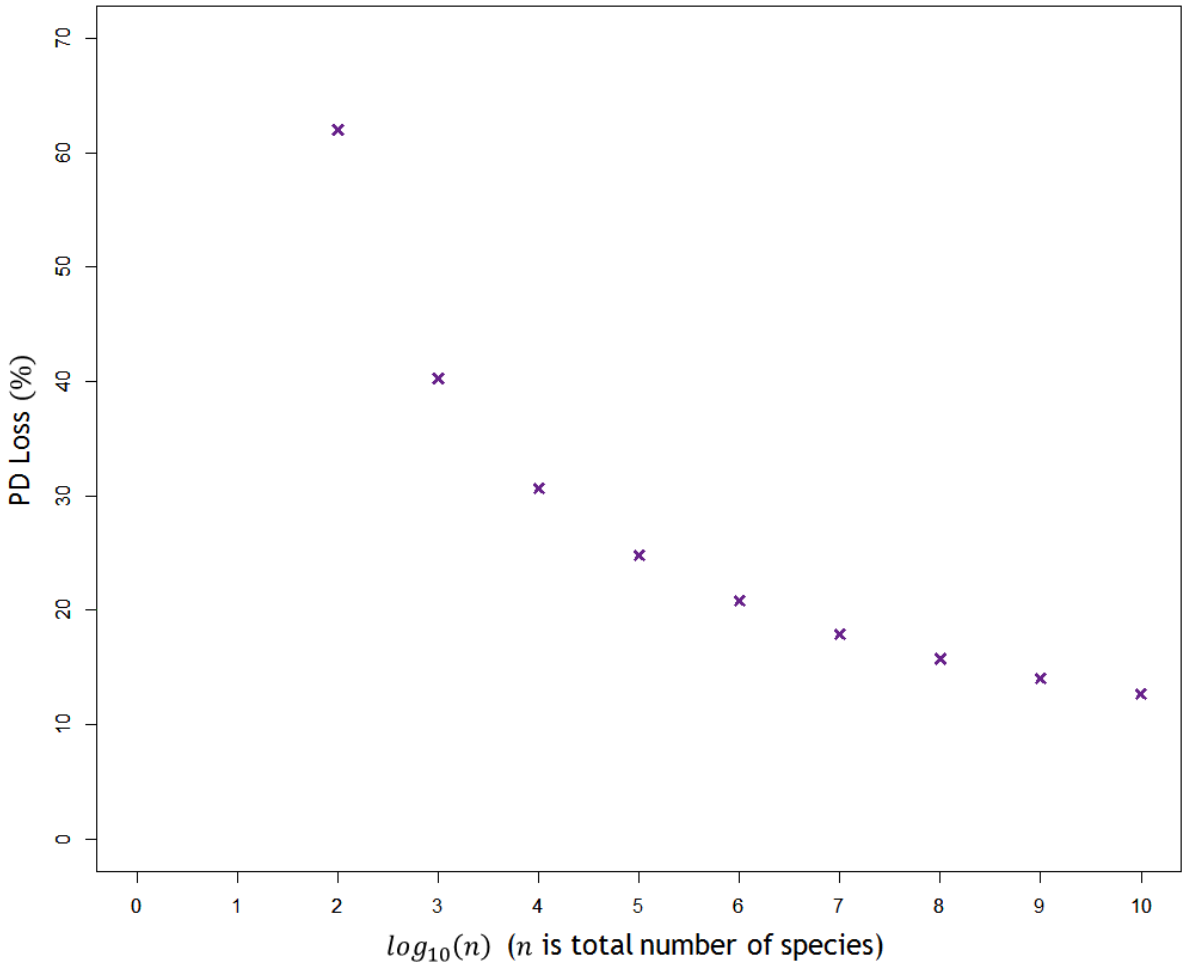

Figure 12: Plot of PD lost using formula (5) when 95% of species are lost.

It is important to note that this equation depends on k **and** n. Figure 12 above shows a plot of the formula, when 95% of species are lost, for a total of 100 species up to 10^{10} species, on a $\log_{10}$ scale. It demonstrates the dependency on n. We look at 95% species loss because we are interested in the claims made in the two papers (section 3.1). We can see that Nee et al.'s claim is true when the total number of species is greater than 10^6, in fact, the total number of species is 5×10^6, according to their research. The formula does not agree with Steel's claim.

This formula suggests that if we were to lose all species existing on Earth at present ($\approx$ 13 million), then only 17.67% of phylogenetic diversity would be lost. This idea is astonishing and rather unbelievable. One would expect that the loss of 95% of all species would lead to a much more significant loss of PD, and devastating effects.

3.3.3 The Use of Different Models

In each paper, two different models are used to simulate speciation trees. Steel et al. use the simple birth model to derive their results, whereas Nee et al. use the coalescent model. This is why the two papers produce two contradicting results. As discussed in section 2.5, the simple birth tree has longer edges which get smaller with each speciation event, meanwhile the coalescent tree has lots of short edges, particularly close to present time. This is why estimated PD loss is much smaller in Nee et al's paper. Edges near the tips of the tree are most likely to be lost for any given proportion of $\frac{k}{n}$, and with the coalescent model, these edges are particularly small.

3.4 Simulations

3.4.1 Functions

To replicate the findings of the two papers, the creation of some functions in R is necessary. We need functions that run the algorithms described in section 2.3.1 and 2.4.4, and also a function that simulates species loss using the Field of Bullets Model. With these functions, we can perform many iterations for different percentages of species loss.

First, we introduce a new column, `lost`, to our data frame described in section 1.2.5 and this is set to `FALSE` for all rows. Our species loss function has two arguments, a data frame representing a species tree and k, the number of species we wish to lose.

The function works in the following way:

1. Sample k species from our n extant species, by uniformly at random selecting k of the rows with `extant` = `TRUE`.
2. Set the sampled rows' `lost` entries to `TRUE`.
3. Work up from the bottom of the data frame, checking for cases where both descendants of a parent species have been lost; if this is true, then their `lost` entry is also set to `TRUE`.
4. Finally, check the rows, starting from the top, for rows where both `child` rows are lost, their `lost` entry is also set to `TRUE`. This continues until a row is found where at least one `child` entry is not lost and this row corresponds to the MRCA of all surviving species.

3.4.2 Comparison with `ape`

Before running simulations to compare the findings of the two papers, it is important to check that the two algorithms are working correctly. We can compare them to functions within the library `ape` (see section 2.3.2) . Figure 13 below demonstrates that the algorithm described in section 2.4.4 works. The simulation using our algorithm matches the simulation using the function from `ape`. We can see that the green lines and red lines are very similar. The histograms in Figure 14 also confirm that the algorithm and `rcoal` (`ape`'s function for simulating coalescent trees) produce similar simulation results.

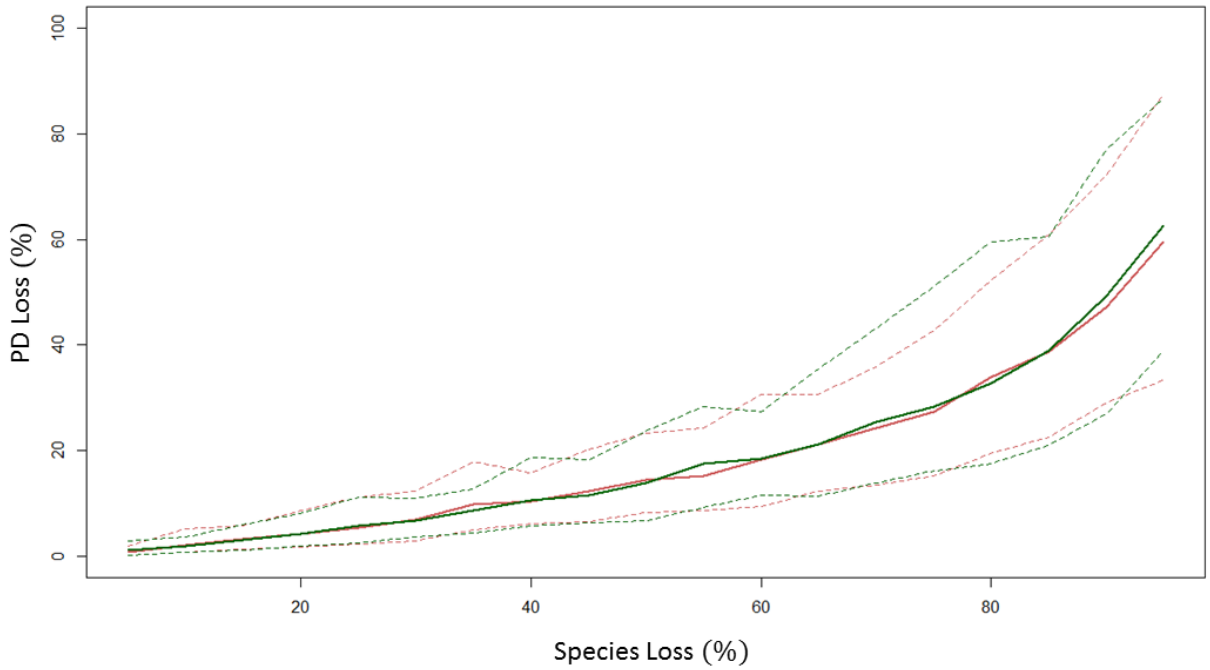

Figure 13: Simulation (red) with 100 species and 100 iterations, using the algorithm for the coalescent model. The green lines represent simulations using `ape`'s function `rcoal`. The solid lines represents the means of PD lost. The dashed lines represent the 95% confidence intervals. The two simulations' means seem to be almost identical and their confidence intervals match relatively well too.

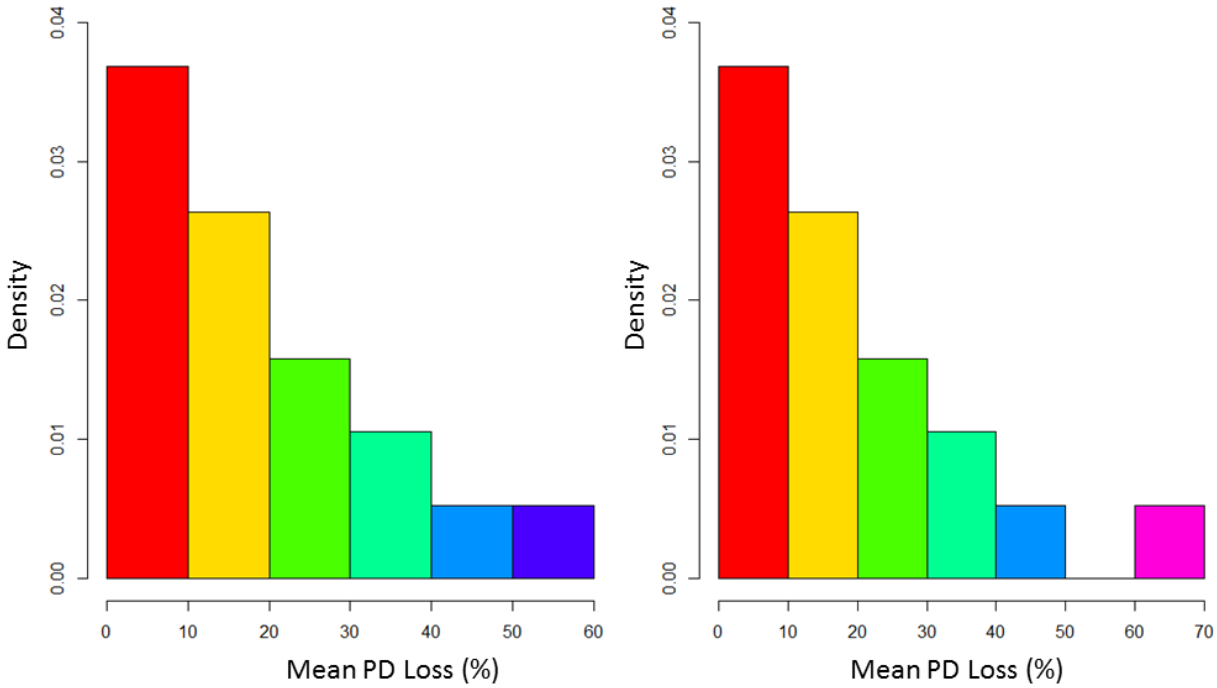

Figure 14: Histograms for the mean PD loss with 100 species and 100 iterations. Simulations using the algorithm (left) vs `ape`'s function `rcoal` (right). The two simulations appear to follow the same distribution since the shapes of the two histograms are very similar.

3.4.3 Simulation Results

Now we shall look at the simulation results. We shall compare the two different models with each other and the approximation formula from the paper by Nee et al. Figure 15 below shows simulations using the birth model and the coalescent model with 100 extant species in total. Figure 16 shows the same simulations but with 1000 extant species. In both plots, we look at species loss from 5% to 95%, in increments of 5%.

From these two graphs, we can see that the mean PD loss for the birth model stays consistent regardless of n, the total number of species in the tree. For the birth model, the confidence interval is narrower for the simulation where $n = 1000$, which we would expect to happen, since confidence intervals heavily depend on sample size. As n increases the simulations for the birth model will move further away from the formula derived by Nee et al. This is due to the formula's dependency on n. Since PD loss decreases as n increases in the formula (as discussed in section 3.3.2), it is natural that the birth model disagrees with it (the formula). Clearly PD loss is a lot greater for the birth model in both graphs.

We can see that for the coalescent model, the simulation matches the formula for $n = 100$ and $n = 1000$, supporting Nee. et al's findings. Unlike the birth model, the mean PD

loss decreases as n increases. The confidence intervals are smaller for $n = 1000$, as we would expect. However for both $n = 100$ and $n = 1000$, the confidence intervals are wider for the coalescent model, compared to the confidence intervals for the birth model. Since the distribution of tree shape is the same, this difference arises from the difference in distribution of edge lengths (see section 2.5). The coalescent model has a greater variance in interevent time than the birth model, so this might be the cause.

For 95% species loss, we can expect to lose about 85% for the birth model for both $n = 100$ and $n = 1000$. This agrees with Steel's findings, suggesting serious implications if 95% of our species on Earth were lost today, as we would naturally expect.

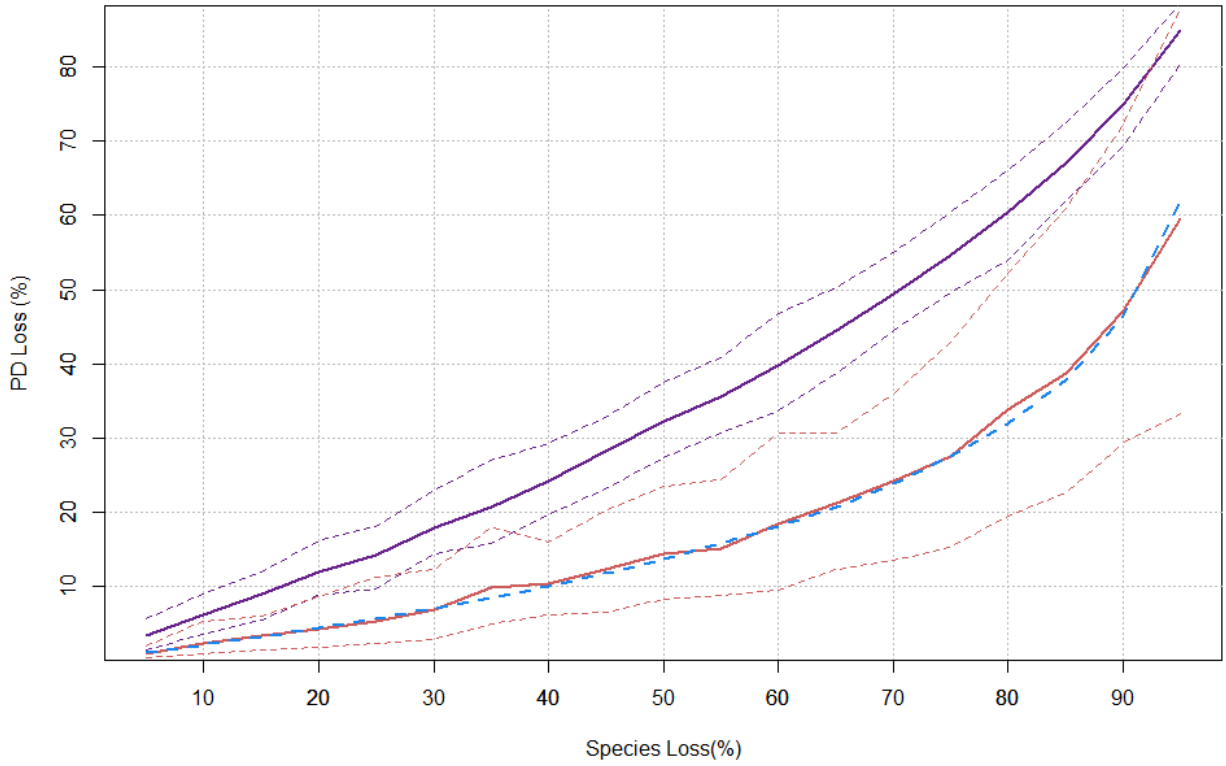

Figure 15: Simulation using the birth model (purple) and coalescent model (red) with $n = 100$ species, in total, and 100 iterations for each species loss value. The solid line represents the mean PD lost. The dashed lines represent the 95% confidence intervals. The blue dashed line is the formula from Nee et al's paper.

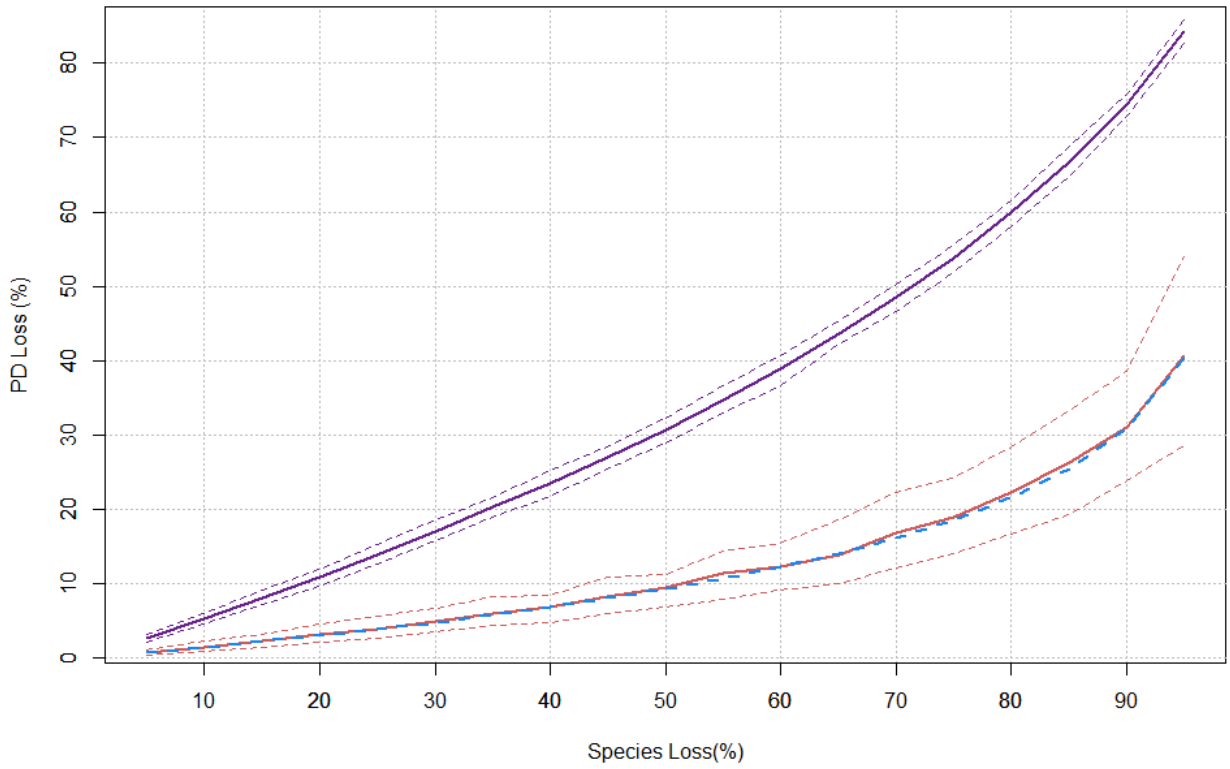

Figure 16: Simulation using the birth model (purple) and coalescent model (red) with $n = 1000$ species, in total, and 100 iterations for each species loss value. The solid line represents the mean PD lost. The dashed lines represent the 95% confidence intervals. The blue dashed line is the formula from Nee et al's paper..

3.5 Further Considerations

Perhaps a model where speciation rate is time-dependent or lineage-dependent would be more appropriate. One must question, how realistic is it to assume that speciation remains constant over time? Surely, changes in the environment have an effect on speciation rate. An extreme change in a species' habitat may cause an increase in speciation, whereas if an ecological equilibrium is reached, there are less external pressures for a given species to adapt. This implies that speciation rates are not constant over time.

4 Inferring Speciation and Extinction Rates

Recall that for the speciation and extinction trees based on the general birth-death model (section 2.3) we have speciation and extinction rates, dependent on the constants λ and μ and the number of extant species. There is freedom to choose the values of λ and μ. However, when studying actual lineages and their evolutionary history, we do not know the values of these constants. Harvey et al. show that it is possible to estimate both λ and μ by using reconstructed phylogenies [6]. We shall look at how this estimation of the birth-death parameters is possible even without 'observing' extinct species - a surprising result!

4.1 Counting Lineages

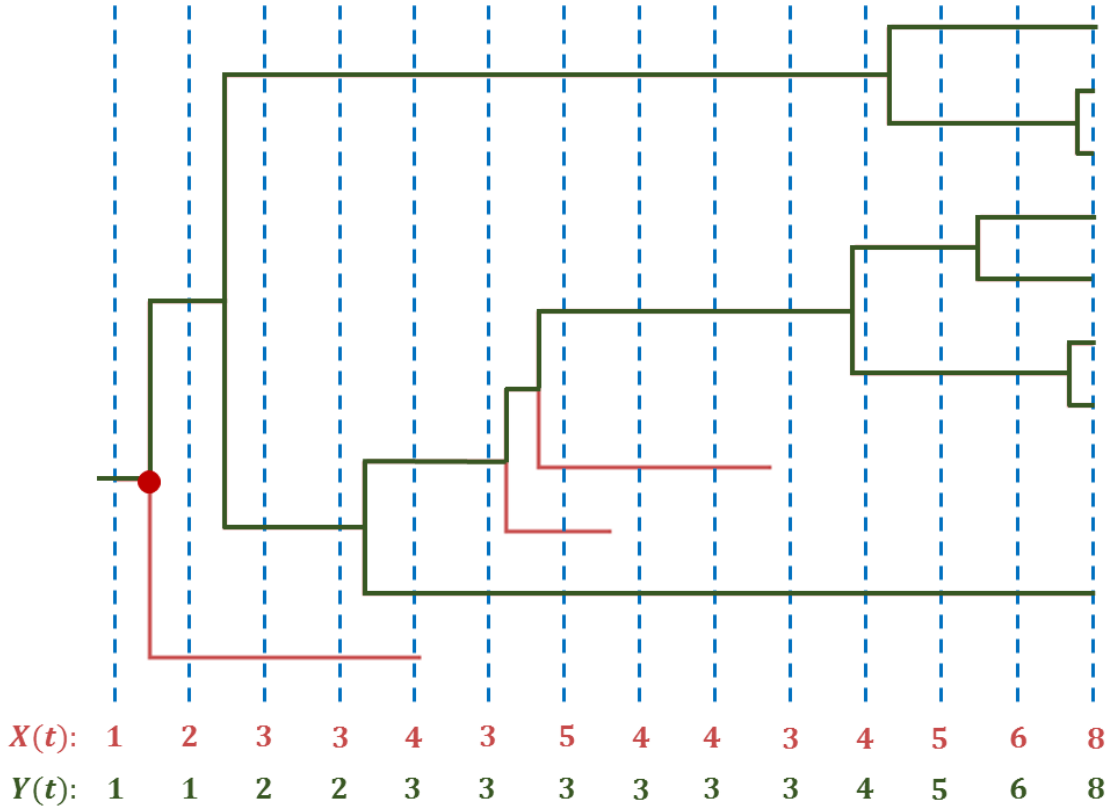

Figure 17: A hypothetical phylogeny with eight extant species. The blue dashed lines mark points in time from the MRCA to present day. The tree represented by the green lines corresponds to the speciation tree based on the extant species. The tree represented by both the green and red lines corresponds to the speciation-extinction tree based on both extant and extinct species. Below the plot are the values of $X(t)$ (red) and $Y(t)$ (green) at each time point. One can see that $Y(t) \leq X(t)$ for all time points t.

In this section we are interested in counting lineages for speciation-extinction trees. Let $X(t)$ denote the number of lineages existing at time t. $X(t)$ follows a birth-death process (see section 2.3) with birth rate λ and death rate μ, starting with $X(0) = 1$ (the MRCA). Let $Y(t)$ denote the number of lineages at time t which have descendant species alive at the present, t_{now}. We have $Y(t) \leq X(t)$, i.e., $Y(t)$ is usually an underestimation of the number of lineages. This is because the lineages that became extinct are not counted, as there is no present day evidence of them. Figure 17 above demonstrates the underestimation. We can see that $X(t)$ is represented by a speciation-extinction tree, whereas $Y(t)$ is represented by the corresponding speciation tree . If we assume there is at least one species alive at the present, then the following hold:

$$Y(0) = 1 \text{ and}$$
$$Y(t) \to X(t), \text{ as } t \to t_{\text{now}}.$$

Figure 18 shows a plot of $X(t)$ and $Y(t)$ for a realisation of a birth-death process with $\lambda = 0.1$ and $\mu = 0.05$, and 1000 extant species. Figure 19 shows a realisation with the same number of extant species, the same λ value and $\mu = 0.075$. These both show that the curve for $Y(t)$ is below the curve for $X(t)$. To produce these plots we create a data frame for 1000 extant species, with the given parameter values, using the algorithm described in section 2.3.1. Two new R functions are required: The first counts the number of species alive at each time in the data frame; the second 'removes' the extinct lineages by deleting any rows corresponding to extinct species or rows which have lost both descendants. This is similar to the species loss function in 3.4.1.

If $X(t)$ and $Y(t)$ are plotted, the line for $Y(t)$ is always below the line for $X(t)$ but meets it at two different points, $t = 0$ and $t = t_{\text{now}}$. From this, we arrive at a surprising result. By considering only the species which survive to t_{now}, we can estimate both λ and μ from these graphs of $Y(t)$. Or more precisely, we can infer speciation and extinction rates from looking at the gradient of a curve plotted for $\text{E}[Y(t)]$, when the number of lineages' axis is on a logarithmic scale. A derivation of this idea will be given below. So, we can potentially infer something about the death rate μ, despite never 'seeing' any extinctions, i.e., having no direct evidence about the number of extinction events or when they occurred.

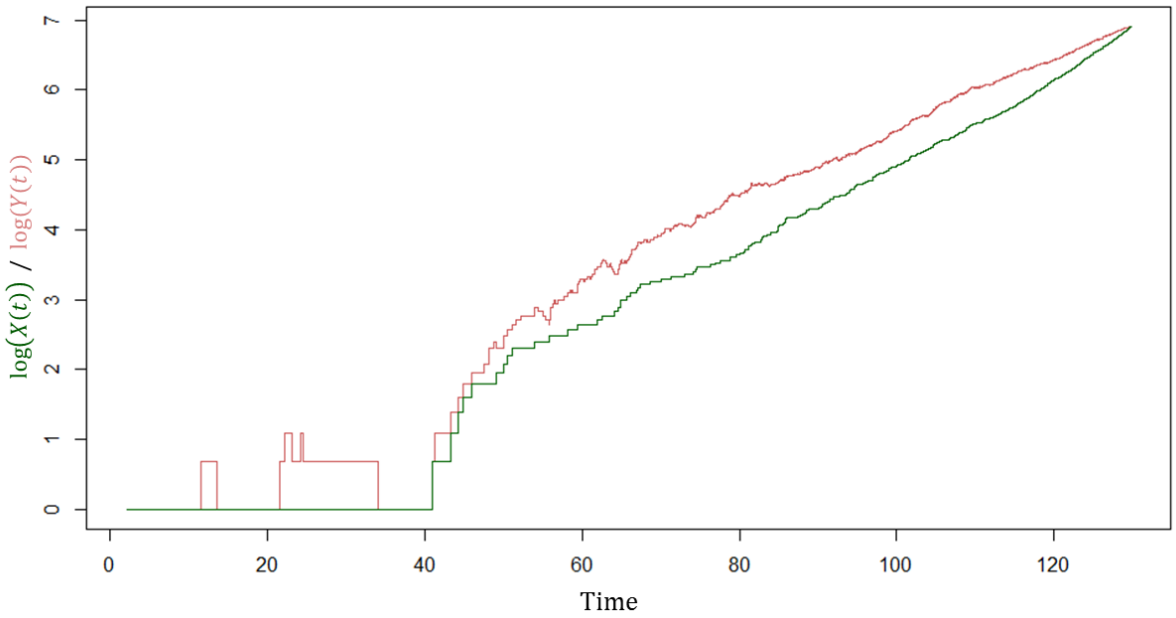

Figure 18: One realisation of a birth-death process with $\lambda = 0.1$ and $\mu = 0.05$.

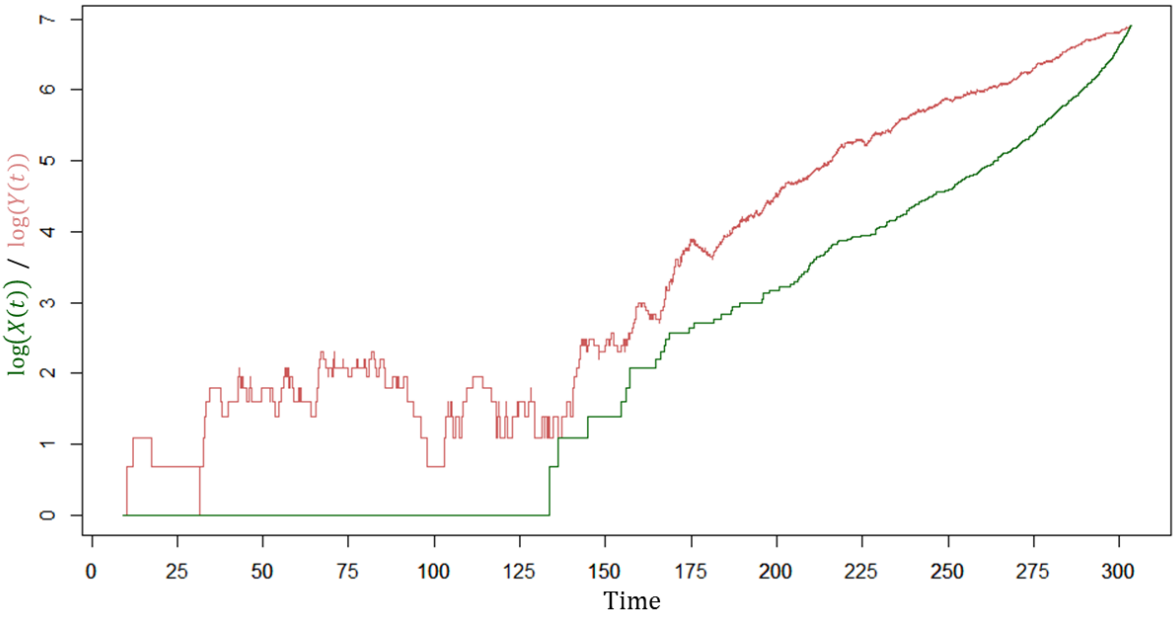

Figure 19: One realisation of a birth-death process with $\lambda = 0.1$ and $\mu = 0.075$.

4.2 Estimation of λ and μ

4.2.1 The Expected Number for $Y(t)$

To understand how we can estimate λ and μ we look at a result from Harvey et al.'s paper. They give an equation for the expected number for $Y(t)$, given $Y(t_{now}) > 0$.

$$\mathrm{E}[Y(t)|Y(t_{now} > 0)] = \mathrm{E}[X(t)] \times \frac{\mathrm{P}(t_{\mathrm{now}} - t)}{\mathrm{P}(t_{\mathrm{now}})}, \tag{6}$$

where $\mathrm{P}(t_{\mathrm{now}} - t)$ is the probability a species at time t has descendants alive at t_{now}. It is worth noting that $\dfrac{\mathrm{P}(t_{\mathrm{now}} - t)}{\mathrm{P}(t_{\mathrm{now}})} \geq 1$, implying that $\mathrm{E}[Y(t)|Y(t_{now} > 0)] \geq \mathrm{E}[X(t)]$. Intuitively this make sense because we expect the number of species alive at time t_{now} to be greater given that total extinction has not occurred, i.e., we know $Y(t) > 0$. We can think of the $\dfrac{\mathrm{P}(t_{\mathrm{now}} - t)}{\mathrm{P}(t_{\mathrm{now}})}$ term as 'pushing' up the expected number and it is has more influence the closer t is to t_{now}. Figure 20 below demonstrates this.

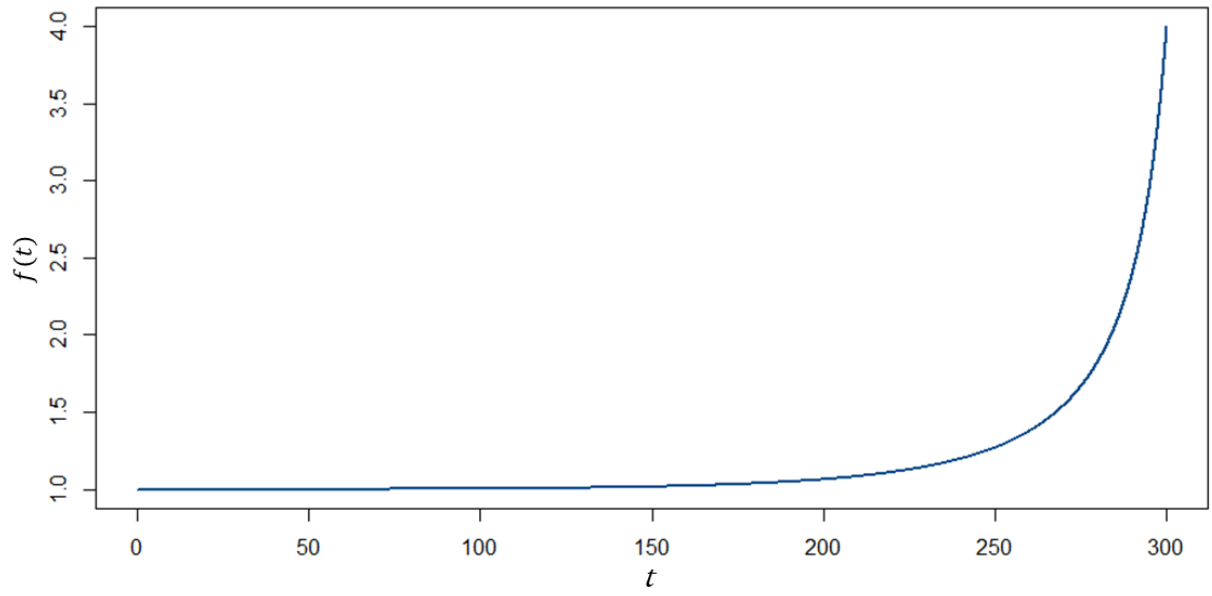

Figure 20: A plot of t against $f(t)$, where $f(t) = \frac{\mathrm{P}(t_{\mathrm{now}}-t)}{\mathrm{P}(t_{\mathrm{now}})}$, with $\lambda = 0.1, \mu = 0.075$ and $t_{\mathrm{now}} = 300$. We can see that as t approaches t_{now}, $f(t)$ increases rapidly away from 1.

From Kendall [8] we have,

$$\mathrm{P}(t_{\mathrm{now}} - t) = \frac{\lambda - \mu}{\lambda - \mu e^{-(\lambda-\mu)(t_{\mathrm{now}}-t)}}. \tag{7}$$

We have $\mathrm{E}[X(t)] = e^{(\lambda-\mu)t}$, from standard birth-death theory.

4.2.2 Finding the Gradient of $\mathrm{E}[Y(t)]$

We shall use the standard method of differentiation to find the gradient for $\mathrm{E}[Y(t)]$. It is convenient to take logs first, so we have

$$\log \mathrm{E}[Y(t)|Y(t_{\text{now}}) > 0] = (\lambda - \mu)t + \log \mathrm{P}(t_{\text{now}} - t) - \log \mathrm{P}(t_{\text{now}}). \tag{8}$$

Noting that with respect to t, the last term is a constant, we differentiate (8) to get the gradient:

$$\frac{d}{dt} \log \mathrm{E}[Y(t)|Y(t_{\text{now}}) > 0] = (\lambda - \mu) + \frac{d}{dt} \log \mathrm{P}(t_{\text{now}} - t).$$

However, $\log \mathrm{P}(t_{\text{now}} - t) = \log(\lambda - \mu) - \log\left[\lambda - \mu e^{-(\lambda-\mu)(t_{\text{now}}-t)}\right]$, so

$$\begin{aligned}
\frac{d}{dt} \log \mathrm{P}(t_{\text{now}} - t) &= \frac{-1}{\lambda - \mu e^{-(\lambda-\mu)(t_{\text{now}}-t)}} \times \frac{d}{dt}\left[\lambda - \mu e^{-(\lambda-\mu)(t_{\text{now}}-t)}\right] \\
&= \frac{\mu}{\lambda - \mu e^{-(\lambda-\mu)(t_{\text{now}}-t)}} \times e^{-(\lambda-\mu)(t_{\text{now}}-t)} \times (\lambda - \mu).
\end{aligned}$$

Recall, from section 4.1, the curves for $X(t)$ and $Y(t)$ meet each other at present day and also at some time far in the past. This is because we do not know exactly when the MRCA first arose. So we consider $\frac{d}{dt} \log \mathrm{P}(t_{\text{now}} - t)$ for two cases: (a) as $t \to t_{\text{now}}$ and (b) as $(t_{\text{now}} - t) \to \infty$. This results in the following:

$$\begin{aligned}
&\text{(a) } \frac{d}{dt} \log \mathrm{P}(t_{\text{now}} - t) \to \mu \text{ and} \\
&\text{(b) } \frac{d}{dt} \log \mathrm{P}(t_{\text{now}} - t) \to 0 \text{ (assuming } \lambda > \mu).
\end{aligned}$$

Hence we have

$$\frac{d}{dt} \log \mathrm{E}[Y(t)|Y(t_{\text{now}}) > 0] \to \lambda, \text{ as } t \to t_{\text{now}}, \text{ and} \tag{9}$$

$$\frac{d}{dt} \log \mathrm{E}[Y(t)|Y(t_{\text{now}}) > 0] \approx \lambda - \mu, \text{ as } (t_{\text{now}} - t) \to \infty. \tag{10}$$

Thus, we have shown that we can infer λ and μ from a plot of $\log \mathrm{E}[Y(t)|Y(t_{\text{now}}) > 0]$ over time t, by calculating the gradient at two different points of the curve. To actually carry out this estimation, in practice, is difficult. From one tree, we will not obtain $\mathrm{E}[Y(t)|Y(t_{\text{now}}) > 0]$ but only a single realisation of $Y(t)$ for $t \in [0, t_{\text{now}}]$. To get a value for $\mathrm{E}[Y(t)|Y(t_{\text{now}}) > 0]$ we need many realisations and then the average number of lineages at set time points. More details on the estimation and its derivation can be found in [6].

5 Conclusion

Speciation-extinction trees are a useful mechanism to map out the branching pattern of speciation and extinction over time. Here I have written R code that can simulate these trees using a data frame structure developed for this project and the R `ape` package. The simple birth model and coalescent model are both used as a basis for the simulation of speciation trees, the case when we do not consider extinction, as in reality we never truly 'witness' extinction events. These simulated trees can be utilised to look at how species loss affects loss of phylogenetic diversity (PD), the total evolutionary history contained within a tree.

An altercation between two papers, both concerned with PD loss, has arisen and after analysis we can see that this is because of the use of different models for speciation. Both papers use the same model, the Field of Bullets model, for species loss, but in slightly different ways. The Field of Bullets model says that extinction uniformly occurs at random and all species have the same probability of becoming extinct. However, how plausible is this model? Is species loss really dictated this way? Therefore, we can question the applicability of the results of the papers. A future consideration would be to look for another model. In fact, other research does exist, for example, maximising algorithms and other methods for conserving species have been explored [4, 9, 14].

In this project there has been a strong focus on 'known speciation trees'. Realistically we do not know the exact branching pattern for all extant species, as we do not have fossil records for all species that have become extinct. We have seen, in the final section of the report, that we can estimate both birth λ and death μ rates from plots of the expected number of lineages in speciation trees. By calculating the gradient of the curves at two different time points we can find $\lambda - \mu$ and λ, and thus we can deduce μ. This astonishing result makes it possible for us to know more about the evolutionary history of the species we see alive today.

References

[1] Allen, L.J.S. (2011), *An Introduction to Stochastic Processes with Applications to Biology; Second Edition*, Chapman and Hall/CRC.

[2] Darwin, C. and Endersby, J. (2009), *On the Origin of Species*, Cambridge University Press.

[3] Dirzo, R., Young, H. S., Galetti, M., Ceballos, G., Isaac, N.J.B. and Collen, B. (2014), 'Defaunation in the Anthropocene', *Science*, 345: 401 - 406.

[4] Erwin, T.L. (1991), 'An Evolutionary Basis for Conservation Strategies', *Science*, 253: 750 - 752.

[5] Felsenstein, J. (2004), *Inferring Phylogenies*, Sinauer Associates.

[6] Harvey, P.H., May, R.M. and Nee, S. (1994), 'Phylogenies without Fossils', *Evolution; International Journal of Organic Evolution*, 48:523-526.

[7] Hein, J., Schierup, M.H. and Wiuf. C. (2005), *Gene Genealogies, Variation and Evolution; A Primer in Coalescent Theory*, Oxford University Press.

[8] Kendall. D.G. (1948), 'On Some Modes of Population Growth Leading to R.A. Fisher's Logarithmic Series Distribution', *Biometrika*, 25: 6 - 15.

[9] Nee, S., May, R.M. (1997), 'Extinction and the Loss of Evolutionary History', *Science*, 278: 692 - 694.

[10] Owen, N. (2014), 'Life on the Edge', *Significance*, 11: 26 - 29.

[11] Page, R.D.M. and Holmes, E.C. (1998), *Molecular Evolution; A Phylogenetic Approach*, Wiley-Blackwell.

[12] Raup, D.M. (1992), *Extinction: Bad Genes or Bad Luck?*, W. W. Norton and Company.

[13] Steel, M. and Lambert, A. (2013), 'Predicting the Loss of Phylogenetic Diversity Under Non-stationary Diversification Models', *Journal of Theoretical Biology*, 337: 111 - 116.

[14] Vane-Wright, R.I., Humphries, C. J. and Williams, P. H.(1991), 'What to Protect? - Systematics and the Agony of Choice', *Biological Conservation*, 55: 235 - 254.

[15] World Wild Fund for Nature (2014), *Living Planet Report 2014: Species and Spaces, People and Places.*