Modeling the Dynamics of Cultural Diversification

1. Diversity and Diversification

Culture can be understood as circulating populations of mental representations (e.g. knowledge, norms, values, beliefs, practices) and their public representations expressed through cultural practices, novel art, organizations, and material culture (e.g. cultural things) (Koch et al, in progress). This tutorial will introduce you to how we can quantify changes in long-lasting representations (e.g. cultural lineages), and broader cultural forms through time.

In this tutorial, you will learn:

- How to think about cultural change in terms of the diversity of cultural lineages
- What we can learn from diversity indices, and how to calculate them
- What we can learn from diversification rates, and how to calculate them
- How to simulate diversification within a population of cultural lineages

a. Quantifying Diversity

Diversity broadly describes the distribution of elements across categories (Leonard and Jones 1989; Stirling 2007). In macroevolutionary biology, these elements are often species and the categories of interest are taxonomic clades, ecological niches, or geographic ranges. In cultural contexts, we are interested in the distribution of mental and material representations. Here, our elements are **cultural lineages** and our categories of interest are **cultural forms** that may include art genres, scientific disciplines, technologies, or institutional structures.

Below we list some diversity measures (adapted from Stirling 2007) that are commonly used across disciplines and what they quantify. The notation below includes the following:

- N is the total number of cultural lineages
- I is a particular cultural lineage
- n_i is the number of i individual things within a cultural lineage

Property	Description	Measurement	
Variety/Richness	Number of lineages	N	How many
Abundance	Count within a lineage	n_i	What is
Distinctiveness	Distance between lineages	Pairwise comparisons on distance metric	How differen

Property	Description	Measurement	
Evenness	Simpson Index	$\sum_i (\frac{n_i}{N})^2$	What is the probability two ra
Evenness	Gini Index	$1 - \sum_i (\frac{n_i}{N})^2$	What is the probability two rand
Evenness	Shannon Entropy	$-\sum_i \frac{n_i}{N} \ln \frac{n_i}{N}$	How evenly distributed ar
Normalized Evenness	Pielou Index	$(-\sum_i \frac{n_i}{N} \ln \frac{n_i}{N}) / \ln N$	How evenly distributed are things ac

The appropriateness of different diversity indices is contingent on the frame of analysis and resolution of the data available. In the populational analyses often conducted in Ecological studies or Anthropological ethnographies, evenness indices work well because we have a fairly reliable measure of the number of individuals in a group or the overall size of a group within a given region. However, over the historial time scales of interest in macroevolution, archaeology, history, and cultural/historical sociology, we may not have reliable estimates of the number things within a lineage. For example, it is difficult to know how many pots were manufactured in a particular style or how many people shared a cultural trait, value, or practice. In these circumstances, variety/richness indices are often most appropriate to use.

The tools we introduce in the following analyses rely on measures of variety/richness or the **"standing diversity"** of cultural lineages. Note that in some contexts where you have abundance data (e.g. the number of tweets belonging to a hashtag), it may be possible to weave these other diversity indices into cultural macroevolutionary analyses.

We note now that even if you don't have complete abundance data, the **methods in these tutorials assume that the long-lasting representations or lineages in your data are either comprise the complete set of the cultural form circulating among the population of people, or minimially a representative sample of this set.** Theoretically, you can't identify the operation of evolutionary mechanisms on cultural diversity without at least a representative sample. Empirically, we have used complete populations of cultural forms whenever possible, such as our analysis of all car models from 1896-2018 or all Metal bands created during the 20th century.

b. From Diversity to Lineages

Diversity indices are useful tools to evaluate and compare the variety, balance, and disparity between different cultural systems. *However, one of the key limitations of diversity indices are that* *they do not disentangle the underlying proceses of cultural innovation and cultural death.* Ideas and products are "born" (i.e. circulate within a population of individuals) when they are publically produced and shared with others. Ideas and products culturally "die" when they stop circulating between people.

c. Diversity in American Automobiles

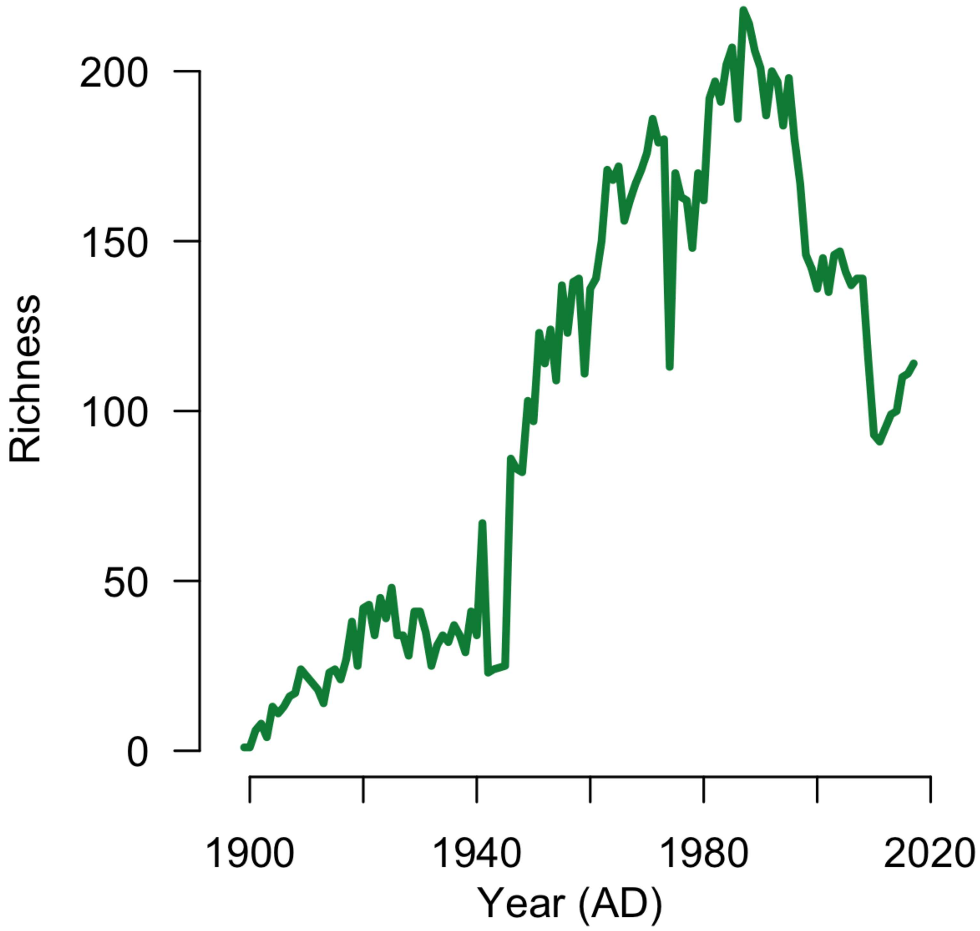
Below we demonstrate some exploratory analyses to look at cultural diversity over time. We feature work on the diversification dynamics of American car models manufactured between 1896 and 2018. We highlight automobiles as they represent one of the most transformative technologies in human history. Since their introduction in the late 19th century, automobiles have radically changed how people move, where people live, and even the global climate. Automobiles are also a diverse technological system (or cultural form) that has gone through many changes but also stayed relatively similar in overall form.

In our analysis, we use car models as an example of cultural lineages. This is because we consider each car model to have a commercial and cultural reality that persists through time despite small generational changes in features and physical apperances. In addition, car models have an established classification/categorization scheme that has emerged throughout the history of the automotive industry. The benefits of using an established classification scheme is that we do not need to quantify all the traits of each car model as would be required for phylogenetic analysis; we only need to know the start and end years of production (occurrence data) for each car model.

The list and production years of each car model derives from the Master Vehicle List maintained by Ebay for the listing and selling of automobile parts. We believe the dataset to be fairly complete and represents a signficant majority of the automobiles manufactured over the last 120 years. The dataset and analysis was originally featured in a recent article (Gjesfjeld et al. 2020) with the associated data available here.

Richness Through Time

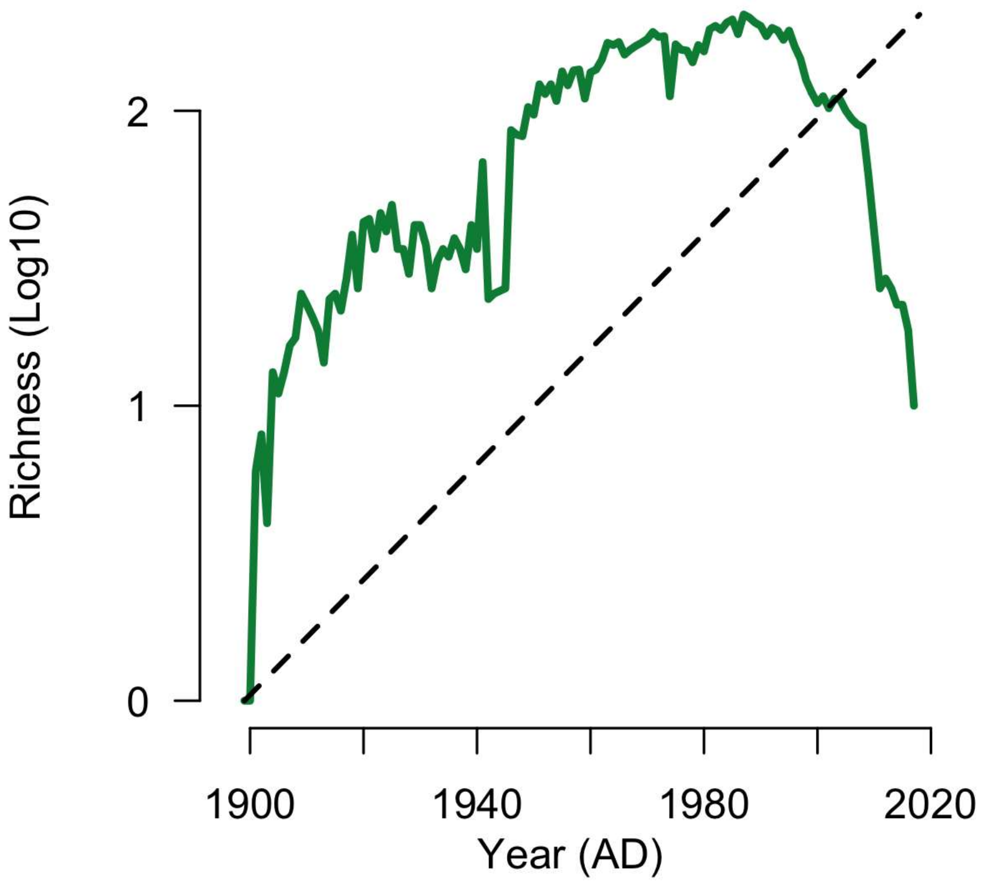
One of the most straightforward plots we can create is the richness or standing diversity through time. More simply, this is the number of car models that are produced each year by American companies. However, since we do not know how many of each car model were made or sold each year, we are unable to calculate the evenness or abundance of car models. A plot of car model richness fom 1896 to 2018 looks like this:



▶ R Code

This plot demonstrates the dramatic increase in the number of different car models produced by American automobile compannies from 1896 to 2018. The general trend is a steady increase until the early 1980s and then a fairly steady decrese in range of car models produced.

An alternative way to view richness through time is the use of a log lineage through time (LTT) plot. This plot uses the common practice of log-transforming the y-axis (or lineage richness). In a plot in semi-logarithmic space, the null expectation is of continued and even growth, which is indicated by a straight line. Deviations from a straight line indicate time periods where the origination or death of new lineages is greater or lower than might be expected from constant growth.

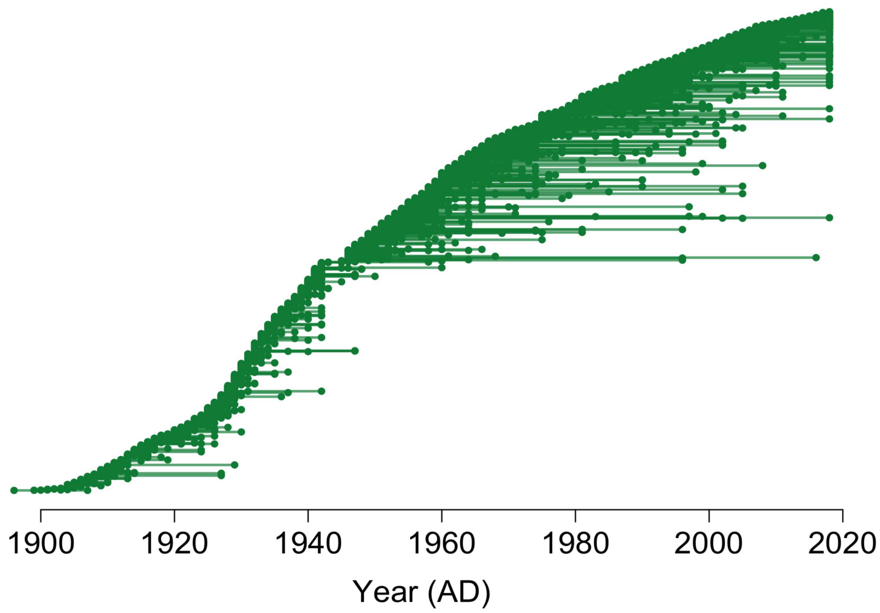


▶ R Code

This LTT plot demonstrates the dramatic increase in the number of different car models produced by American automobile compannies from 1896 to 2018. The general trend is a steady increase until the early 1980s and then a fairly steady decrese in range of car models produced.

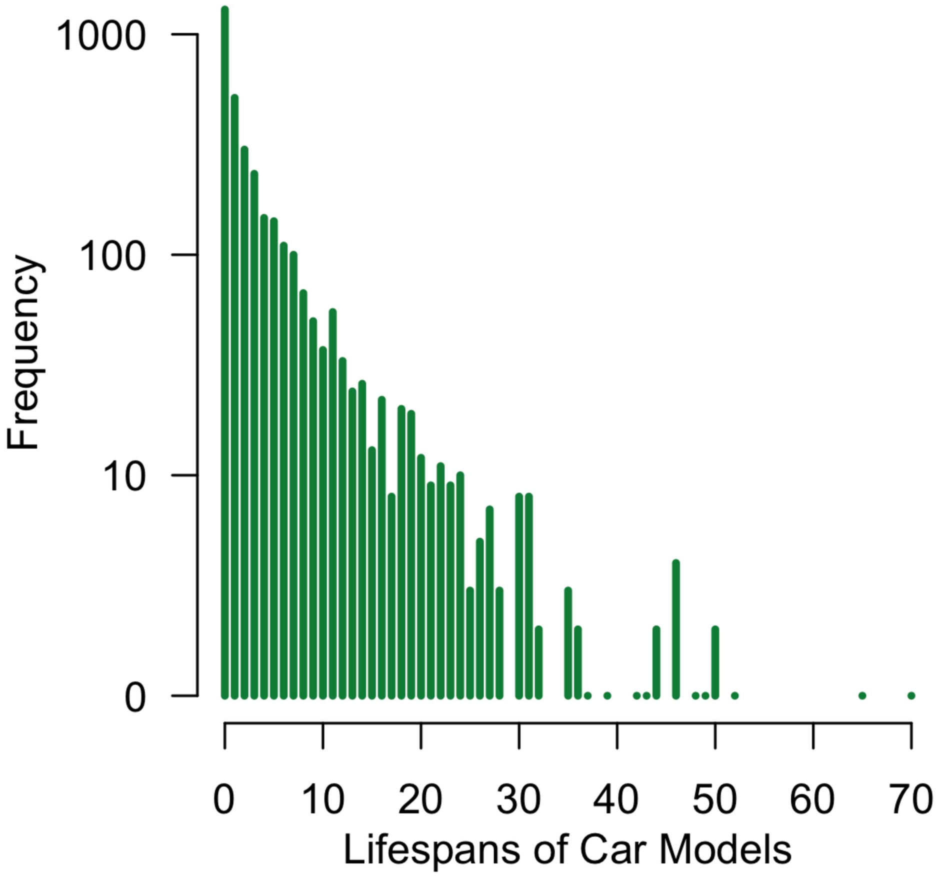
Longevity

In addition to simply plotting changes in richness through time, we can also evaluate the longevity of different car models. As we can see in the plot below, some car models have incredibly long lifespans ("living fossils") while others have very short lifespans. We can also see that shorter lifespans tend to be more prevlant during the early history of automobile production (before 1945), as visualized below.



▶ R Code

We can also create a histogram of the lifespans presented above to view the relative proportion of short-lived and long-live car models. As expected, our dataset contains far more car models that have very short lifespans and only a few models with long lifespans.



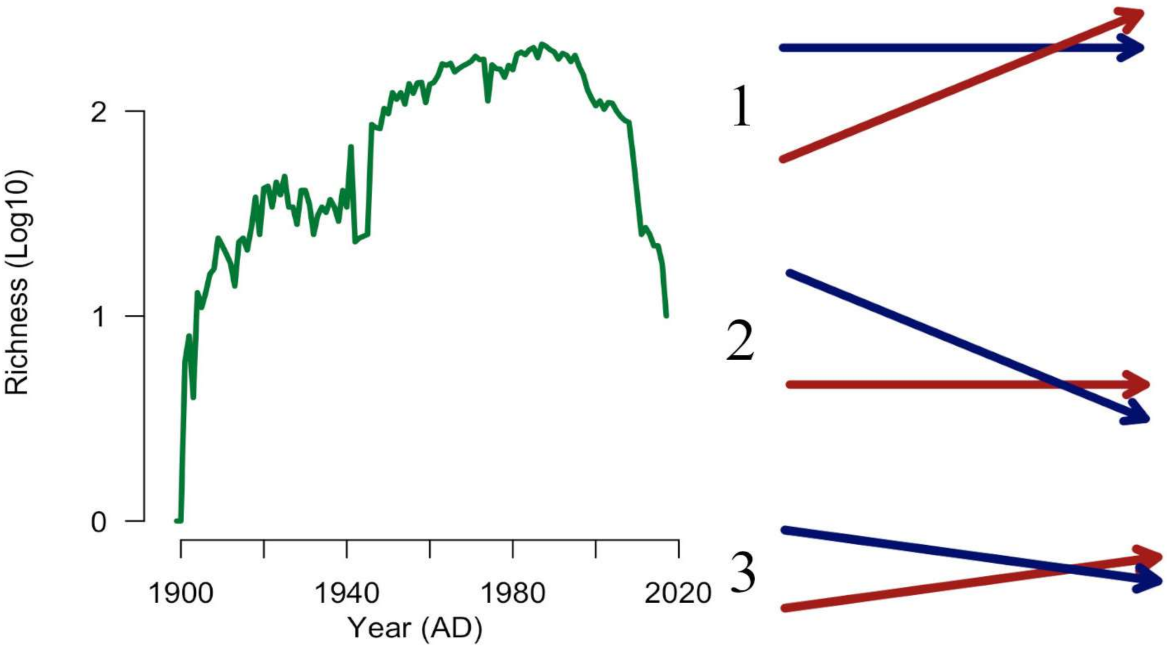
▸ R Code

d. From Diversity Indices to Diversification Rates

Diversity indices are useful tools to evaluate and compare the variety, balance, and disparity within a cultural form. *However, one of the key limitations of diversity indices are that they do not disentangle the underlying proceses of cultural innovation and cultural death.* Thefore, instead of diversity indices, many macroevolutionary approaches (including ours) focus on birth/origination and death/extinction rates as the primary metric of interest. Compared to variety/richness, diversification rates provide additional insight into the dynamics of stability and change over time.

For example in car models, the log lineage through time plot shows a dramatic decline in the diversity of car models from the mid-1990s until today. But this decline could be caused by,

1) A stable orignation rate and rising extinction rate
2) A stable extinction rate and declining origination rate
3) A combination of a declining origination rate and rising extinction rate



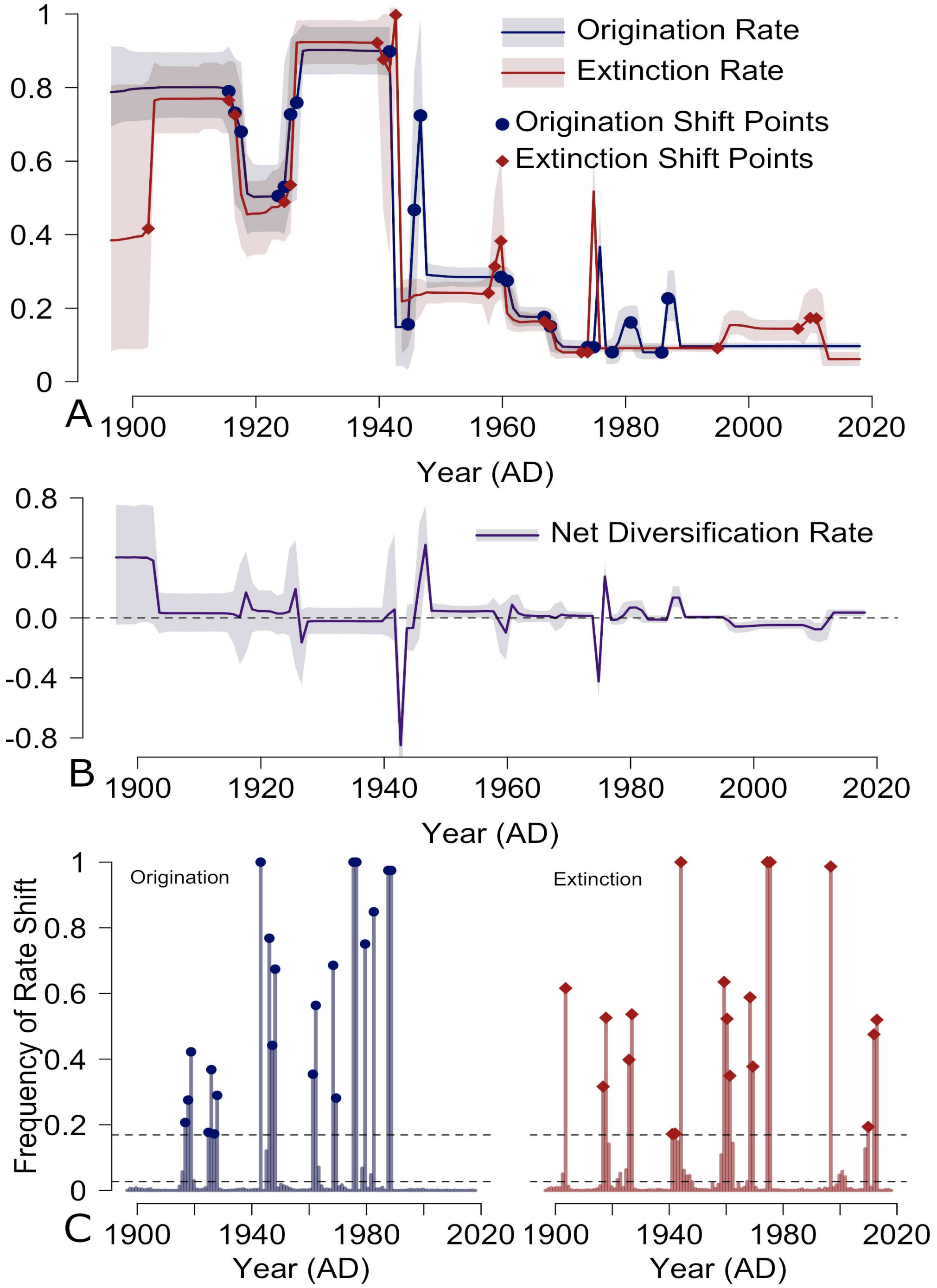
Diversification rates simply provide more information into the processes underlying changes in diversity than diversity indices. For example, one could hypothesize that each of the three diversification scenarios presented above represent a different set of underlying cultural, economic or social processes.

1) The first scenario might represent an economic slowdown where some car models are simply no longer produced, similar to what we saw in the 2008 Recession.

2) The second scenario may suggest a pattern of competition where diversity is slowly being reduced over time due to the emergence of a "dominant design", as has been suggested in economics literature (Abernathy 1978).

3) The third scenario suggests the possible combination of interacting processes, such as an economic slowdown that will raise extinction rates in combination with a uncertain environment that would tend to decrease origination rates.

In any of these circumstances, understanding the relationship between origination and extinction rates is imperative to gaining further insight processes driving the patterns of diversity. This is clear when we view the diversification of American automobiles from 1896-2018, as opposed to simply the change in car model richness over time shown above. By estimating both oringation and extinction rates, we are able to identify historical periods in which both rates were impacted (Great Depression, WWII) and those in which only one was impacted (Arab Oil Crises, 2008 Recession). Futhermore, we can also identify with far greater clarity the years that are estimated to have a signficant shift in either origination and extinction rates, which are highlighted in the figure below from Gjesfjeld et al. (2020)



Why look at diversification rates?

Diversification rates capture how cultural lineages "beget" other lineages through learning and innovation by people, as well as how cultural lineages "die" from disuse and forgetting. They do not make strong assumptions about individual-level circumstances (as in agent-based simulations), or the exact sequence of transmission and variation events (as in networks and phylogenies). Diversification rates consider the entire population or cultural form without privileging some lineages as more important than others. We all have our own heterogeneous personal cultures in which some representations and things are more relevant than others, so in many cases this is a reasonable simplifying assumption. However it is a strong assumption, and should be evaluated in context.

Diversification rate analyses of cultural lineages should be considered as a complement, not a competitor, of actor-based analyses like social network analysis or agent-based simulations. While those methods start from actor-level transmission processes to extrapolate population-scale cultural outcomes, diversification rate analysis take the opposite approach. Diversification rate analysis starts from population-level cultural phenomena and identifies trends that are consistent with individual-level processes occuring. This approach can be helpful in corroborating simulations, or when actor-level structural data simply aren't available.

e. Formal Introduction to Diversification Rates

In macroevolutionary biology, certain patterns in diversification rates are recognized as theoretically consistent with evolutionary mechanisms that have shaped the dynamics of the species or clade over time. In cultural contexts, birth and death rates can highlight the role of major events and evolutionary mechanisms in the histories of cultural forms.

The birth rate is defined as the expected number of birth events per lineage per time unit (e.g. 1 year) and can be approximated by the following formula:

$$\lambda = \frac{number\ of\ lineage\ births}{total\ time\ lived}$$

The death rate is defined as the expected number of death events per lineage per time unit and can be approximated as:

$$\mu = \frac{number\ of\ lineage\ deaths}{total\ time\ lived}$$

where *total time lived* is the total time collectively lived by all lineages in the period of analysis. If the period of analysis is just a single unit of time (e.g., year), this reduces to:

$$\lambda = \frac{number\ of\ lineage\ births}{standing\ diversity}$$

and

$$\mu = \frac{number\ of\ lineage\ deaths}{standing\ diversity}$$

f. Creating a Diversification Rate Simulator

To clarify what diversification rates actually represent, we are going to create a diversification rate simulator.

In this first example, we can simulate population dynamics over time by randomly selecting individual lineages to reproduce or die in each time unit. Remember, the number of lineages in the population at any moment is sometimes called the net or **standing diversity**. If we calculate the raw diversification rates from the standing diversity over time, we call these the **empirical** birth and death rates (as opposed to **estimated or theoretical** rates from statistical models).

Let's start by creating a population object that can birth individuals, kill individuals, and keep track of the standing diversity.

To create this object, hover over the brackets below (next to SHOW CODE) and then click on the "play" button. If you're comfortable programming, you can double-click to get the gist of the code (even if you're not familar with Python). **Just make sure you run the code blocks in sequential order otherwise you will get errors.**

```
#@title
import numpy as np
class Population:

  def __init__(self,starting_diversity,total_time):
    self.total_time=total_time #total number of time units (e.g. years)
    self.birth_times=np.zeros(starting_diversity) #every individual in starting p
    self.death_times=np.repeat(total_time,starting_diversity) #for now, set all i

    self.alive_index=np.arange(starting_diversity)

  def currently_alive(self):
    return self.alive_index

  def create_individuals(self,num_individuals,time):
    '''
    create "num_individuals" at "time"
    '''

    #update alive index
    self.alive_index=np.concatenate( (self.alive_index, np.arange(len(self.birth_
    #update birth times
    self.birth_times=np.concatenate( ( self.birth_times,np.repeat(time, num_indiv
    self.death_times=np.concatenate( ( self.death_times,np.repeat(self.total_time
  def kill_individuals(self,indices,time):
    '''
    kill off individuals with "indices" in alive_index at "time"
    '''
    #update death times
    kill_index=self.alive_index[indices] #need to get back from alive index to ti
    self.death_times[kill_index]=time

    #update alive index
    self.alive_index=np.delete(self.alive_index,indices)

  def calc_time_lived(self, frame_start, frame_end): #found in literate library a
    '''
    calculates total time lived by all individuals within a time window
    '''
    s, e  = self.birth_times.astype(float), self.death_times.astype(float)
    s[s<frame_start] = frame_start #set elements born before timeframe to start o
    e[e>frame_end] = frame_end # set elements dying after timeframe to end of tim
    dt = e - s
    return np.sum(dt[dt>0])
```

Now we will also create an object that simulates population dynamics over time. This object has parameters for all the characteristics we are interested in. These include:

- Time Length: How many time units we want to simulate for
- Starting Diversity: The diversity or richness of lineages at the first time step
- Theoretical Birthrate: The rate at which new lineages emerge in each time step
- Theoretical Deathrate: The rate at which existing linages are lost in each time step
- Random Seed: A random value with which to reproduce the stochastic simulation

Once again, hover over the brackets below and then click on the "play" button.

```
#@title
import pandas as pd
import tqdm
import warnings #to deal with stupid tqdm bug and dep warning for sol
warnings.filterwarnings('ignore')
class Simulator:
  def __init__(self,
    theoretical_lambda,
    theoretical_mu,
    epoch=50,
    starting_div=1000,
    ):
    self.theoretical_lambda=theoretical_lambda
    self.theoretical_mu=theoretical_mu
    self.epoch=epoch
    self.starting_div=starting_div

  def run_simulation(self, seed=None):
    if seed!=None: np.random.seed(seed)
    pop=Population(self.starting_div,self.epoch)

    #storage arrays
    empirical_lambda=[]
    empirical_mu=[]
    standing_diversity=[]

    for t in tqdm.tqdm_notebook(range(1,self.epoch+1)):
      #stochastically birth and kill individuals in each time unit
      r=np.random.sample(len(pop.currently_alive())) #each living individual gets
      birther_indices = (r < self.theoretical_lambda).nonzero()[0] #get indices o
      dying_indices = np.intersect1d((r >= self.theoretical_lambda).nonzero()[0],
      '''
      Breaking this down for R users (where it would be prettier)
      1. r >= birth_rate returns bool array
      2. nonzero returns nonzero indices
      3. intersect1d returns numbers in both arrays
      '''

      time_lived=pop.calc_time_lived(t-1,t)
      assert time_lived == len(pop.currently_alive()) #just to show you

      #update population
      pop.create_individuals(len(birther_indices),t)
      pop.kill_individuals(dying_indices,t)

      #calculate statistics
      empirical_lambda.append(len(birther_indices)/time_lived)
      empirical_mu.append(len(dying_indices)/time_lived)
      standing_diversity.append(time_lived)

    return pd.DataFrame({
        'Time': np.arange(self.epoch),
        'Theoretical Birthrate': self.theoretical_lambda,
        'Theoretical Deathrate': self.theoretical_mu,
        'Empirical Birthrate': empirical_lambda,
        'Empirical Deathrate': empirical_mu,
        'Standing Diversity': standing_diversity
    })
```

g. Simulating constant diversification rates

Note the default settings:

- Time Length: 50
- Starting Diversity: 1000
- Theoretical Birthrate: 0.01
- Theoretical Deathrate: 0.005
- Random Seed: 12345

Before moving on to plot the outcome of the simulation, what would be your initial expectation for how the standing diversity changes through time? Would it increase or decrease? And by how much? Adjust the sliding scales below to match these values if needed, and press the "play" button.

We've made the diversification rates unrealistically small here to minimize major stochastic swings. We'll account for these in more elegant ways in the future, but we want to keep the code simple. Please note that we are primarily interested in the relative differences between birth and death rates rather than their absolute magnitudes.

time_length: 50

starting_diversity: 105

theoretical_birthrate: .01

theoretical_deathrate: .01

random_seed: 111

3. Finally, let's make the birth and death rates equal to each other. Change the time length back to 50 and set both the birth and death rates to 0.1. What does the standing diversity look like with these settings?
   Now, change the random seed from 12345 to 111 and then to 222 and then 333. How much does the standing diversity change inbetween these different settings? **What does this tell us about the stochasticity in our simulation?**

Beyond getting a feel for the relationship between diversification rates and standing diversity, it's important that you notice that **the observed empirical rates are noisy stochastic representations of the true, theoretical underlying rates.** If we are to use rate analyses to understand the histories of cultural forms, we must be able to differentiate between stochastic noise and meaningful changes in the rates over time. This motivates the usage of statistical models for birth and death rates that cut through this noise with some modest assumptions, as we will highlight in future tutorials.

Key Takeaways

- **Culture can be understood as circulating populations of cultural representations, which we refer to as cultural lineages. Changes in the diversity of these cultural lineages can explain how culture emerges, stabilizes, or changes over time.**

- **There are a variety of indices for highlighting different aspects of diversity and its change over time.**

- **Diversification (birth and death) rates contain more information than indices because they describe processes of cultural origination and extinction.**

- **Empirical diversifications rates are calculated as the number of births/deaths over total time lived in a time window. These snapshots are noisy representations of the true/theoretical rates.**

Up Next...

In the next lesson, we introduce you to the linear birth-death process, the core statistical framework for our diversification rate methods and the LiteRate algorithm to estimate statistically-significant rate shifts.

References

Abernathy, William. The Productivity Dilemma: Roadblock to Innovation in the Automobile Industry. Baltimore ; London: Johns Hopkins University Press, 1978.

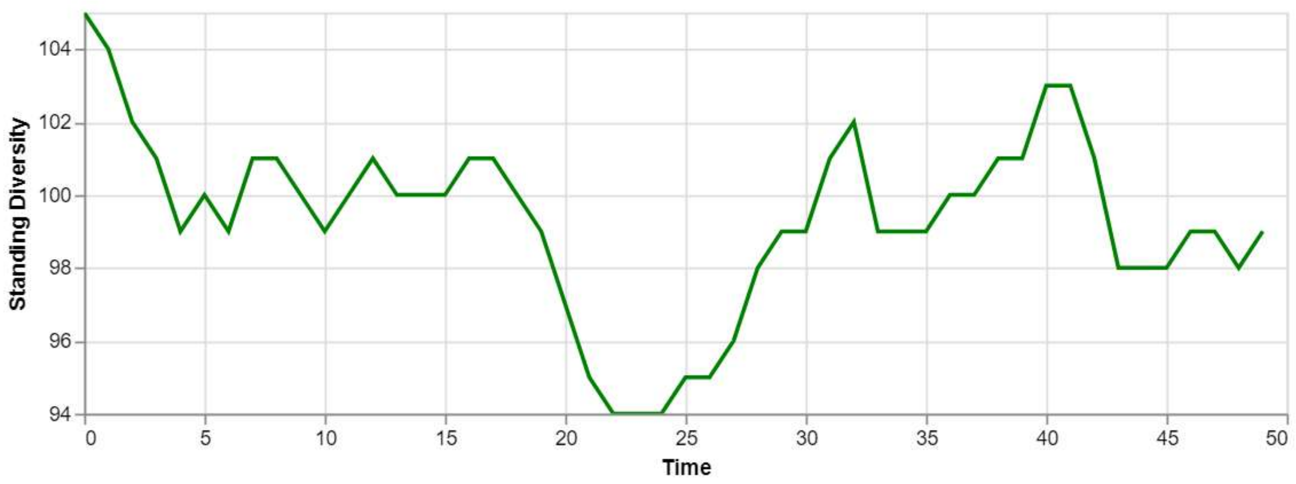
Gjesfjeld, Erik, Daniele Silvestro, Jonathan Chang, Bernard Koch, Jacob G. Foster, and Michael E. Alfaro. ‘A Quantitative Workflow for Modeling Diversification in Material Culture’. PLOS ONE 15, no. 2 (6 February 2020): e0227579. https://doi.org/10.1371/journal.pone.0227579.

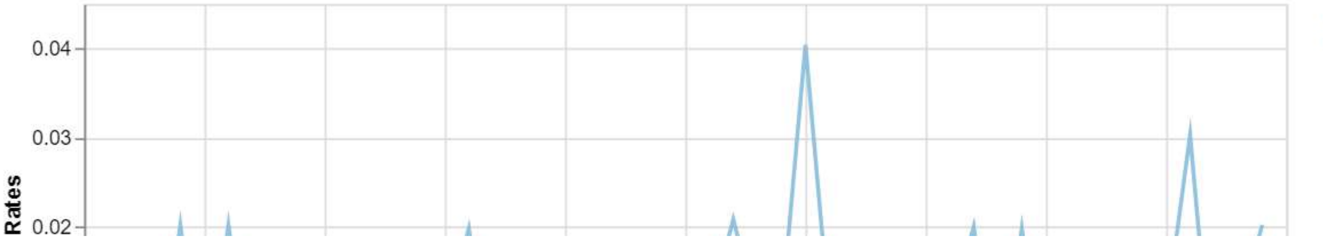
Koch, Bernard, Daniele Silvestro, and Jacob G. Foster. n.d. “The Evolutionary Dynamics of Cultural Change (as Told Through the Birth and Brutal, Blackened Death of Metal Music).” SocArXiv. osf.io/preprints/socarxiv/659bt.

Leonard, Robert D, and George T Jones, eds. Quantifying Diversity in Archaeology. Cambridge University Press Cambridge, 1989.

Stirling, Andy. ‘A General Framework for Analysing Diversity in Science, Technology and Society’. Journal of the Royal Society Interface 4, no. 15 (2007): 707–719. https://doi.org/10.1098/rsif.2007.0213

```
100%                                              50/50 [00:00<00:00, 1229.97it/s]
```





With the settings above, we can see that the standing diversity steadily increases from 1000 lineages to 1300 lineages over 50 time steps. Because we have a theoretical birth rate that is higher than the theoretical death rate, we should expect to see an increase in the standing diversity through time.

However, the empirical birth and death rates show much greater variability from time step to time step. Over long enough periods and with big enough populations, the empirical rates will match the theoretical rates, but as you can see they are quite variable between each time unit.

Check Your Understanding:

Using the simulator above, try to answer the following questions:

1. What happens if you change the theoretical birth rate to a much higher value, such as 0.1? Before you run it, what type of growth do you expect to see (linear, exponential, logistic) in the standing diversity?
   What happens to the empirical rates over time. Why?

2. Keeping the birth rate at 0.1 and the death rate at 0.05, change the time length from 50 to 5. What happens to the standing diversity? What type of growth does it look like now?