Introduction

This book was born of research in category theory, brought to life by the ongoing vigorous debate on how to quantify biological diversity, given strength by information theory, and fed by the ancient field of functional equations. It applies the power of the axiomatic method to a biological problem of pressing concern, but it also presents new advances in 'pure' mathematics that stand in their own right, independently of any application.

The starting point is the connection between diversity and entropy. We will discover:

- how Shannon entropy, originally defined for communications engineering, can also be understood through biological diversity (Chapter 2);
- how deformations of Shannon entropy express a spectrum of viewpoints on the meaning of biodiversity (Chapter 4);
- how these deformations *provably* provide the only reasonable abundance-based measures of diversity (Chapter 7);
- how to derive such results from characterization theorems for the power means, of which we prove several, some new (Chapters 5 and 9).

Complementing the classical techniques of these proofs is a large-scale categorical programme, which has produced both new mathematics and new measures of diversity now used in scientific applications. For example, we will find:

- that many invariants of size from across the breadth of mathematics (including cardinality, volume, surface area, fractional dimension, and both topological and algebraic notions of Euler characteristic) arise from one single invariant, defined in the wide generality of enriched categories (Chapter 6);
- a way of measuring diversity that reflects not only the varying abundances of

species (as is traditional), but also the varying similarities between them, or, more generally, any notion of the values of the species (Chapters 6 and 7);
- that these diversity measures belong to the extended family of measures of size (Chapter 6);
- a 'best of all possible worlds', an abundance distribution on any given set of species that maximizes diversity from an infinite number of viewpoints simultaneously (Chapter 6);
- an extension of Shannon entropy from its classical context of finite sets to distributions on a metric space or a graph (Chapter 6), obtained by translating the similarity-sensitive diversity measures into the language of entropy.

Shannon entropy is a fundamental concept of information theory, but information theory contains many riches besides. We will mine them, discovering:

- how the concept of relative entropy not only touches subjects from Bayesian inference to coding theory to Riemannian geometry, but also provides a way of quantifying local diversity within a larger context (Chapter 3);
- quantitative methods for identifying particularly unusual or atypical parts of an ecological community (Chapter 8, drawing on work of Reeve et al. [293]).

The main narrative thread is modest in its mathematical prerequisites. But we also take advantage of some more specialized bodies of knowledge (large deviation theory, the theory of operads, and the theory of finite fields), establishing:

- how probability theory can be used to solve functional equations (Chapter 9, following work of Aubrun and Nechita [20]);
- a streamlined characterization of information loss, as a natural consequence of categorical and operadic thinking (Chapters 10 and 12);
- that the concept of entropy is (provably) inescapable even in the pure-mathematical heartlands of category theory, algebra and topology, quite separately from its importance in scientific applications (Chapter 12);
- the right definition of entropy for probability distributions whose 'probabilities' are elements of the ring $\mathbb{Z}/p\mathbb{Z}$ of integers modulo a prime p (Chapter 11, drawing on work of Kontsevich [195]).

The question of how to quantify diversity is far more mathematically profound than is generally appreciated. This book makes the case that the theory of diversity measurement is fertile soil for new mathematics, just as much as the neighbouring but far more thoroughly worked field of information theory.

* * *

What *is* the problem of quantifying diversity? Briefly, it is to take a biological community and extract from it a numerical measure of its 'diversity' (whatever that should mean). This task is certainly beset with practical problems: for instance, field ecologists recording woodland animals will probably observe the noisy, the brightly coloured and the gregarious more frequently than the quiet, the camouflaged and the shy. There are also statistical difficulties: if a survey of one community finds 10 different species in a sample of 50 individuals, and a survey of another finds 18 different species in a sample of 100, which is more diverse?

However, we will not be concerned with either the practical or the statistical difficulties. Instead, we will focus on a fundamental conceptual problem: in an ideal world where we have complete, perfect data, how can we quantify diversity in a meaningful and logical way?

In both the news media and the scientific literature, the most common meaning given to the word 'diversity' (or 'biodiversity') is simply the number of species present. Certainly this is an important quantity. However, it is not always very informative. For instance, the number of species of great ape on the planet is 8 (Example 4.3.8), but 99.99% of all great apes belong to just one species: us. In terms of global ecology, it is arguably more accurate to say that there is effectively only one species of great ape.

An example illustrates the spectrum of possible interpretations of the concept of diversity. Consider two bird communities:

A

B

In community A, there are four species, but the majority of individuals belong to a single dominant species. Community B contains the first three species in equal abundance, but the fourth is absent. Which community, A or B, is more diverse?

One viewpoint is that the presence of *species* is what matters. Rare species count for as much as common ones: every species is precious. From this view-

point, community A is more diverse, simply because more species are present. The abundances of species are irrelevant; presence or absence is all that matters.

But there is an opposing viewpoint that prioritizes the balance of *communities*. Common species are important; they are the ones that exert the most influence on the community. Community A has a single very common species, which has largely outcompeted the others, whereas community B has three common species, evenly balanced. From this viewpoint, community B is more diverse.

These two viewpoints are the two ends of a continuum. More precisely, there is a continuous one-parameter family $(D_q)_{q\in[0,\infty]}$ of diversity measures encoding this spectrum of viewpoints. Low values of q attach high importance to rare species; for example, D_0 measures community A as more diverse than community B. When q is high, D_q is most strongly influenced by the balance of more common species; thus, D_∞ judges B to be more diverse. No single viewpoint is right or wrong. Different scientists adopt different viewpoints (that is, different values of q) for different purposes, as the literature amply attests (Examples 4.3.5).

Long ago, it was realized that the concept of diversity is closely related to the concept of entropy. Entropy appears in dozens of guises across dozens of branches of science, of which thermodynamics is probably the most famous. (The introduction to Chapter 2 gives a long but highly incomplete list.) The most simple incarnation is Shannon entropy, which is a real number associated with any probability distribution on a finite set. It is, in fact, the logarithm of the diversity measure D_1. Most often, Shannon entropy is explained and understood through the theory of coding; indeed, we provide such an explanation here. But the diversity interpretation provides a new perspective.

For example, the diversity measures D_q, known in ecology as the Hill numbers, are the exponentials of what information theorists know as the Rényi entropies. From the very beginning of information theory, an important role has been played by characterization theorems: results stating that any measure (of information, say) satisfying a list of desirable properties must be of a particular form (a scalar multiple of Shannon entropy, say). But what counts as a desirable property depends on one's perspective. We will prove that the Hill numbers D_q are, in a precise sense, the only measures of diversity with certain natural properties (Theorem 7.4.3). This theorem translates into a new characterization of the Rényi entropies, but it is not one that necessarily would have been thought of from a purely information-theoretic perspective.

However, something is missing. In the real world, diversity is understood as involving not only the number and abundances of the species, but also how *dif-*

ferent they are. (For example, this affects conservation policy; see the OECD quotation on p. 169.) We describe the remedy in Chapter 6, defining a family of diversity measures that take account of the varying similarity between species, while still incorporating the spectrum of viewpoints discussed above. This definition unifies into one family a large number of the diversity measures proposed and used in the ecological and genetics literature.

This family of diversity measures first appeared in a paper in *Ecology* [220], but it can also be understood and motivated from a purely mathematical perspective. The classical Rényi entropies are a family of real numbers assigned to any probability distribution on a finite *set*. By factoring in the differences or distances between points (species), we extend this to a family of real numbers assigned to any probability distribution on a finite *metric space*. In the extreme case where $d(x, y) = \infty$ for all distinct points x and y, we recover the Rényi entropies. In this way, the similarity-sensitive diversity measures extend the definition of Rényi entropy from sets to metric spaces.

Different values of the viewpoint parameter $q \in [0, \infty]$ produce different judgements on which of two distributions is the more diverse. But it turns out that for any metric space (or in biological terms, any set of species), there is a single distribution that maximizes diversity from all viewpoints simultaneously. For a generic finite metric space, this maximizing distribution is unique. Thus, almost every finite metric space carries a canonical probability distribution (not usually uniform). The maximum diversity itself is also independent of q, and is therefore a numerical invariant of metric spaces. This invariant has geometric significance in its own right (Section 6.5).

We go further. One might wish to evaluate an ecological community in a way that takes into account some notion of the values of the species (such as phylogenetic distinctiveness). Again, there is a sensible family of measures that does this job, extending not only the similarity-sensitive diversity measures just described, but also further measures already existing in the ecological literature. The word 'sensible' can be made precise: as soon as we subject an abstract measure of the value of a community to some basic logical requirements, it is forced to belong to a certain one-parameter family (σ_q) (Theorem 7.3.4), which are essentially the Rényi *relative* entropies.

Information theory also helps us to analyse the diversity of metacommunities, that is, ecological communities made up of a number of smaller communities such as geographical regions. The established notions of relative entropy, conditional entropy and mutual information provide meaningful measures of the structure of a metacommunity (Chapter 8). But we will do more than simply translate information theory into ecological language. For example, the new characterization of the Rényi entropies mentioned above is a byproduct of

the characterization theorem for measures of ecological value. In this way, the theory of diversity gives back to information theory.

* * *

The scientific importance of biological diversity goes far beyond the obvious setting of conservation of animals and plants. Certainly such conservation efforts are important, and the need for meaningful measures of diversity is well appreciated in that context. For example, Vane-Wright et al. [342] wrote thirty years ago of the 'agony of choice' in conservation of flora and fauna, and emphasized how crucial it is to use the right diversity measures.

But most life is microscopic. Nee [262] argued in 2004 that

> [w]e are still at the very beginning of a golden age of biodiversity discovery, driven largely by the advances in molecular biology and a new open-mindedness about where life might be found,

and that

> all of the marvels in biodiversity's new bestiary are invisible.

Even excluding exotic new discoveries of microscopic life, two recent lines of research illustrate important uses of diversity measures at the microbial level.

First, the extensive use of antimicrobial drugs on animals unfortunate enough to be born into the modern meat industry is commonly held to be a cause of antimicrobial resistance in pathogens affecting humans. However, a 2012 study of Mather et al. [246] suggests that the causality may be more complex. By analysing the diversity of antimicrobial resistance in *Salmonella* taken from animal populations on the one hand, and from human populations on the other, the authors concluded that the animal population is 'unlikely to be the major source of resistance' for humans, and that 'current policy emphasis on restricting antimicrobial use in domestic animals may be overly simplistic'. The diversity measures used in this analysis were the Hill numbers D_q mentioned above and central to this book.

Second, the increasing problem of obesity in humans has prompted research into causes and treatments, and there is evidence of a negative correlation between obesity and diversity of the gut microbiome (Turnbaugh et al. [335, 336]). Almost all traditional measures of diversity rely on a division of organisms into species or other taxonomic groups, but in this case, only a fraction of the microbial species concerned have been isolated and classified taxonomically. Researchers in this field therefore use DNA sequence data, applying sophisticated but somewhat arbitrary clustering algorithms to create artificial species-like groups ('operational taxonomic units'). On the other hand,

the similarity-sensitive diversity measures mentioned above and introduced in Chapter 6 can be applied directly to the sequence data, bypassing the clustering step and producing a measure of genetic diversity. A test case was carried out in Leinster and Cobbold [220] (Example 4), with results that supported the conclusions of Turnbaugh et al.

Despite the wide variety of uses of diversity measures in biology, none of the mathematics presented in this text is intrinsically biological. Indeed, the mathematics of diversity was being developed as early as 1912 by the economist Corrado Gini [118] (best known for the Gini coefficient of disparity of wealth), and by the statistician Udny Yule in the 1940s for the analysis of lexical diversity in literature [361]. Some of the diversity measures most common in ecology have recently been used to analyse the ethnic and sociological diversity of judges (Barton and Moran [30]), and the similarity-sensitive diversity measures that are the subject of Chapter 6 have been used not only in multiple ecological contexts (as listed after Example 6.1.8), but also in non-biological applications such as computer network security (Wang et al. [347]).

In mathematical terms, simple diversity measures such as the Hill numbers are invariants of a probability distribution on a finite set. The similarity-sensitive diversity measures are defined for any probability distribution on a finite set with an assigned degree of similarity between each pair of points. (This includes any finite metric space or graph.) The value measures are defined for any finite set equipped with a probability distribution and an assignment of a nonnegative value to each element. The metacommunity measures are defined for any probability distribution on the cartesian product of a pair of finite sets. Much of this text is written using ecological terminology, but the mathematics is entirely general.

* * *

This work grew out of a general category-theoretic study of size. In many parts of mathematics, there is a canonical notion of the size of the objects of study: sets have cardinality, vector spaces have dimension, subsets of Euclidean space have volume, topological spaces have Euler characteristic, and so on. Typically, such measures of size satisfy analogues of the elementary inclusion-exclusion and multiplicativity formulas for counting finite sets:

$$|X \cup Y| = |X| + |Y| - |X \cap Y|,$$
$$|X \times Y| = |X| \cdot |Y|.$$

(The interpretation of Euler characteristic as the topological analogue of cardinality is not as well known as it should be; this is an insight of Schanuel on

which we elaborate in Section 6.4.) From a categorical perspective, it is natural to seek a single invariant unifying all of these measures of size.

Some unification is achieved by defining a notion of size for categories themselves, called *magnitude* or Euler characteristic. (Finiteness hypotheses are required, but will not be mentioned in this overview.) This definition already brings together several established invariants of size [210]: cardinality of sets, and the various notions of Euler characteristic for partially ordered sets, topological spaces, and even orbifolds (whose Euler characteristics are in general not integers). The theory of magnitude of categories is closely related to the theory of Möbius–Rota inversion for partially ordered sets [301, 215].

But the decisive, unifying step is the generalization of the definition of magnitude from categories to the wider class of *enriched* categories [216], which includes not only categories themselves, but also metric spaces, graphs, and the additive categories that are a staple of homological algebra.

The definition of the magnitude of an enriched category unifies still more established invariants of size. For example, in the representation theory of associative algebras, one frequently considers the indecomposable projective modules, which form an additive category. The magnitude of that additive category turns out to be the Euler form of a certain canonical module, defined as an alternating sum of dimensions of Ext groups (equation (6.20)). Magnitude for enriched categories can also be realized as the Euler characteristic of a certain Hochschild-like homology theory of enriched categories, in the same sense that the Jones polynomial for knots is the Euler characteristic of Khovanov homology [189]. This was established in recent work led by Shulman [224], building on the case of magnitude homology for graphs previously developed by Hepworth and Willerton [144].

Since any metric space can be regarded as an enriched category, the general definition of the magnitude of an enriched category gives, in particular, a definition of the magnitude $|X| \in \mathbb{R}$ of a metric space X. Unlike the other special cases just mentioned, this invariant is essentially new.

Recent, increasingly sophisticated, work in analysis has connected magnitude with classical invariants of geometric measure. For example, for a compact subset $X \subseteq \mathbb{R}^n$ satisfying certain regularity conditions, if one is given the magnitude of all of the rescalings tX of X (for $t > 0$), then one can recover:

- the Minkowski dimension of X (one of the principal notions of fractional dimension), a result proved by Meckes using results in potential theory (Theorem 6.5.9);
- the volume of X, a result proved by Barceló and Carbery using PDE methods (Theorem 6.5.6);

- the surface area of X, a result proved by Gimperlein and Goffeng using global analysis (or more specifically, tools for computing heat trace asymptotics; see Theorem 6.5.8).

Gimperlein and Goffeng also proved an asymptotic inclusion-exclusion principle:

$$|t(X \cup Y)| + |t(X \cap Y)| - |tX| - |tY| \to 0$$

as $t \to \infty$, for sufficiently regular $X, Y \subseteq \mathbb{R}^n$ (Section 6.5). This is another manifestation of the cardinality-like nature of magnitude.

We have seen that every finite metric space X has an unambiguous maximum diversity $D_{\max}(X) \in \mathbb{R}$, defined in terms of the similarity-sensitive diversity measures (p. 5). We have also seen that X has a magnitude $|X| \in \mathbb{R}$. These two real numbers are not in general equal (ultimately because probabilities or species abundances are forbidden to be negative), but they are closely related. Indeed, $D_{\max}(X)$ is always equal to the magnitude of some *subspace* of X, and in important families of cases is equal to the magnitude of X itself. So, magnitude is closely related to maximum diversity. Indeed, this relationship was exploited by Meckes to prove the result on Minkowski dimension.

There is a historical surprise. Although this author arrived at the definition of the magnitude of a metric space by the route of enriched category theory, it had already arisen in earlier work on the quantification of biodiversity. In 1994, the environmental scientists Andrew Solow and Stephen Polasky carried out a probabilistic analysis of the benefits of high biodiversity ([319], Section 4), and isolated a particular quantity that they called the 'effective number of species'. They did not investigate it mathematically, merely remarking mildly that it 'has some appealing properties'. It is exactly our magnitude.

* * *

Ecologists began to propose quantitative definitions of biological diversity in the mid-twentieth century [314, 351], setting in motion more than sixty years of heated debate, with dozens of further proposed diversity measures, hundreds of scholarly papers, at least one book devoted to the subject [240], and consequently, for some, despair (expressed as early as 1971 in a famously titled paper of Hurlbert [150]). Meanwhile, parallel debates were taking place in genetics and other disciplines.

The connections between diversity measurement on the one hand, and information theory and category theory on the other, are fruitful for both mathematics and biology. But any measure of biological diversity must be justifiable in purely biological terms, rather than by borrowing authority from information

theory, category theory, or any other field. The ecologist E. C. Pielou warned against attaching ecological significance to diversity measures for anything other than ecological reasons:

> It should not be (but it is) necessary to emphasize that the object of calculating indices of diversity is to solve, not to create, problems. The indices are merely numbers, useful in some circumstances but not in all. [...] Indices should be calculated for the light (not the shadow) they cast on genuine ecological problems.

(Pielou [283], p. 293).

In a series of incisive papers beginning in 2006, the conservationist and botanist Lou Jost insisted that whatever diversity measures one uses, they must exhibit *logical behaviour* [166, 167, 168, 169]. For example, Shannon entropy is commonly used as a diversity measure by practising ecologists, and it does behave logically if one is only using it to ask whether one community is more or less diverse than another. But as Jost observed, any attempt to reason about percentage changes in diversity using Shannon entropy runs into logical absurdities: Examples 2.4.7 and 2.4.11 describe the plague that exterminates 90% of species but only causes a 17% drop in 'diversity', and the oil drilling that simultaneously destroys *and* preserves 83% of the 'diversity' of an ecosystem. It is, in fact, the *exponential* of Shannon entropy that should be used for this purpose.

In this sense, origin stories are irrelevant. Inventing new diversity measures is easy, and it is nearly as easy to tell a story of how a new measure fits with some intuitive idea of diversity, or to justify it in terms of its importance in some related discipline. But if a measure does not pass basic logical tests (as in Section 4.4), it is useless or worse.

Jost noted that all of the Hill numbers D_q do behave logically. Again, we go further: Theorem 7.4.3 states that the Hill numbers are in fact the *only* measures of diversity satisfying certain logically fundamental properties. (At least, this is so for the simple model of a community in terms of species abundances only.) This is the ideal of the axiomatic approach: to prove results stating that if one wishes to have a measure with such-and-such properties, then it can only be one of *these* measures.

Mathematically, such results belong to the field of functional equations. We review a small corner of this vast and classical theory, beginning with the fact that the only measurable functions $f : \mathbb{R} \to \mathbb{R}$ satisfying the Cauchy functional equation $f(x + y) = f(x) + f(y)$ are the linear mappings $x \mapsto cx$. Building on classical results, we obtain new axiomatic characterizations of a variety of measures of diversity, entropy and value. We also explain a new method, pio-