6-2021

Unifying Community Detection Across Scales from Genomes to Landscapes

Stephanie F. Hudon
Boise State University

Andrii Zaiats
Boise State University

Anna Roser
Boise State University

Anand Roopsind
Boise State University

Cristina Barber
Boise State University

See next page for additional authors

Publication Information

Hudon, Stephanie F.; Zaiats, Andrii; Roser, Anna; Roopsind, Anand; Barber, Cristina; Robb, Brecken C.; . . . and Caughlin, T. Trevor. (2021). "Unifying Community Detection Across Scales from Genomes to Landscapes". *Oikos, 130*(6), 831-843. https://doi.org/10.1111/oik.08393

Unifying Community Detection Across Scales from Genomes to Landscapes

Stephanie F. Hudon[†]*****
Boise State University,
Boise, ID, USA
stephaniehudon@boisestate.edu

Andrii Zaiats[†]
Boise State University
Boise, ID, USA

Anna Roser
Boise State University
Boise, ID, USA

Anand Roopsind
Boise State University
Boise, ID, USA

and

Center for Natural Climate Solutions
Conservation International
Arlington, VA, USA

Cristina Barber
Boise State University
Boise, ID, USA

Brecken C. Robb
Boise State University
Boise, ID, USA

Britt A. Pendleton
Boise State University
Boise, ID, USA

Meghan J. Camp
Washington State University
Pullman, WA, USA

Patrick E. Clark
Northwest Watershed Research Center
USDA Agricultural Research Service
Boise, ID, USA

Merry M. Davidson
Boise State University
Boise, ID, USA

Jonas Frankel-Bricker
Boise State University
Boise, ID, USA

Marcella Fremgen-Tarantino
Boise State University
Boise, ID, USA

Jennifer Sorensen Forbey
Boise State University
Boise, ID, USA

Eric J. Hayden
Boise State University
Boise, ID, USA

Lora A. Richards
University of Nevada, Reno
Reno, NV, USA

Olivia K. Rodriguez
Boise State University
Boise, ID, USA

T. Trevor Caughlin*
Boise State University
Boise, ID, USA
trevorcaughlin@boisestate.edu

* Corresponding authors.

[†] These authors contributed equally to this work.

Abstract

Biodiversity science encompasses multiple disciplines and biological scales from molecules to landscapes. Nevertheless, biodiversity data are often analyzed separately with discipline-specific methodologies, constraining resulting inferences to a single scale. To overcome this, we present a topic modeling framework to analyze community composition in cross-disciplinary datasets, including those generated from metagenomics, metabolomics, field ecology, and remote sensing. Using topic models, we demonstrate how community detection in different datasets can inform the conservation of interacting plants and herbivores. We show how topic models can identify members of molecular, organismal, and landscape-level communities that relate to wildlife health, from gut microbes to forage quality. We conclude with a future vision for how topic modeling can be used to design cross-scale studies that promote a holistic approach to detect, monitor, and manage biodiversity.

Keywords: biodiversity, metagenomics, metabolomics, Latent Dirichlet Allocation, wildlife conservation, sagebrush

Introduction

Understanding biodiversity will require crossing disciplinary boundaries to link biological organization across scales. Early efforts to quantify biodiversity focused on the organismal scale of plants and animals (Simpson 1949). However, modern biodiversity research encompasses molecular scales, as well as scales beyond individual organisms, including biotic and abiotic features within landscapes, regions, and continents. Studying biodiversity at microscopic and macrosystem scales has led to insights with relevance for human health (Mohajeri *et al.* 2018), global sustainability (Bennett *et al.* 2015) and wildlife conservation (Trevelline *et al.* 2019). As recognition of the importance of biodiversity has increased, so have methods for analyzing biodiversity, from molecular approaches such as Next Generation sequencing for genomic data to airborne sensors that can measure large-scale landscape features. These discipline-specific methods limit analysis of biodiversity patterns that may be nested within or interact among scales. The lack of interdisciplinary cohesion in biodiversity studies with different terminology and varying scales of interest is a barrier to understanding biological processes vital to biodiversity conservation. One step toward overcoming this lack of cohesion is to identify patterns in data across disciplines that can then be discussed with a common language (Mosher *et al.* 2020).

Community organization is a unifying pattern in biodiversity data across scales. Ecological communities of species that occupy the same space at the same time are a major focus of empirical and theoretical work in community ecology (Vellend 2010). More recently, ideas from community ecology have been extended to other biological disciplines, including detecting co-occurring microbes (Nemergut *et al.* 2013), functional genes (Burke *et al.* 2011), and landscape features (Räsänen *et al.* 2016). Despite differences in community assembly and study techniques, co-occurring features can reveal ecologically meaningful patterns in metabolites, microbial taxa, plant and animal species, and spectral bands from land surface reflectance. Community detection across multiple scales opens the possibility to study cross-scale interrelationships. For example, metabolite features within plants influence the microbial organisms of individual herbivores (Kohl *et al.* 2014b) and reflectance features of plants can predict herbivore population dynamics across landscapes (Fauchald *et al.* 2017). The overarching importance of communities in ecology and evolution has led to a multitude of methods to detect communities in ecological data (Legendre and Legendre 2012).

A common challenge of detecting communities is mixed membership, when single features and single samples can potentially be assigned to more than one community. The degree of mixed membership in communities depends on whether features arrange themselves as discrete members of different communities (Clements 1936), or as fluid entities with membership in multiple communities (Gleason 1926). Within cellular units, biomolecular processes such as mutation and differential gene expression can promote mixing of metabolic and genetic features. Within a landscape, processes such as dispersal and anthropogenic disturbances lead to mixing of species and obscure boundaries between communities (Lortie *et al.* 2004). Another challenge for community detection is the tradeoff between sampling extent and resolution, a methodological choice that can affect community membership results. For example, in metagenomics, the benefit of deep sequencing must be weighed against the cost of generating more reads. Similarly, in remote sensing, larger pixels capture more surface area than smaller pixels, but the higher resolution of smaller pixels improves detection of land cover features (Kennedy *et al.* 2009).

Altogether, mixed membership of features within sampling units and communities is common. Nevertheless, many analytical methods, such as clustering and ordination techniques (McCune *et al.* 2002), lack a probabilistic interpretation of community membership, which limits the potential for model transferability and prediction in novel environments. One solution is topic modeling of community membership, which has revolutionized multivariate analysis by enabling a single feature or sampling unit to belong to multiple communities. The term "topic model" arises from text mining, where models are used to assign co-occurring words in documents to underlying subjects ("topics"; Barde and Bainwad 2017). Topics are referred to as "latent" because they are not known before hand and must be inferred from the data. Latent Dirichlet Allocation (LDA) is a topic modeling approach that can identify communities of features, while allowing for mixed membership of features across communities as well as mixtures of communities within individual sampling units (Valle *et al.* 2014). LDA was first developed in population genetics, motivated by the need to use genotypes as features that could group individuals into populations, while allowing for admixture (i.e., the presence of several distinct genotypes/genomes in a single population) (Pritchard *et al.* 2000). Several years later, LDA was independently developed as a tool to uncover latent structure in text data and broadly adopted by the machine learning community (Blei *et al.* 2003). Since then, LDA has resulted in transformative biological insights across disciplines including annotating unknown chemicals in fermented beverages (van Der Hooft *et al.* 2016), characterizing functional roles of gene regions (Chen *et al.* 2010) and identifying communities of bird species in citizen science data (Valle *et al.* 2018). Beyond single discipline applications, we contend that topic modeling has unrealized potential to unify biodiversity science across scales.

Here we demonstrate how to apply LDA across multiple scales to inform conservation of herbivores. We focus on the sagebrush steppe ecosystem that once covered ~ 1 million km^2 of land in the western United States but is increasingly threatened by wildfires and invasive species (Requena-Mullor *et al.* 2019). Sagebrush (*Artemisia* spp.) are the dominant plant species in these ecosystems and are critical for two sagebrush obligate species: the pygmy rabbit (*Brachylagus idahoensis*) and the Greater sage-grouse (*Centrocercus urophasianus*, hereafter sage-grouse). Both herbivores are considered species of conservation concern across the Intermountain West. However, efforts to conserve and reintroduce populations of pygmy rabbits and sage-grouse have had mixed success due to problems that range from lack of consideration of local diet adaptations (Oh *et al.* 2019) to ecosystem fragmentation (Cross *et al.* 2018).

Management of threatened species, including pygmy rabbits and sage-grouse, will benefit from a deeper and more functional understanding of the biological communities that impact individual health and population dynamics. We use four case studies from the sagebrush steppe ecosystem to show how LDA can assess community mixtures of (1) metabolites from leaf material of individual sagebrush plants, (2) microbial species from fecal pellets of pygmy rabbits, (3) plant species from field plots within sagebrush patches, and (4) spectra from pixels across a sagebrush landscape (Figure 1). At the micro-scale, microbial features in herbivores (Kohl *et al.* 2016) interact with metabolite concentrations in the gut after herbivores consume sagebrush (White *et al.* 1982). At the macro-scale, features of herbivores and sagebrush are dependent on metabolite concentrations of plant taxa within habitat patches (Ulappa *et al.* 2014, Frye et al. 2013) and those plant taxa can be detected with aerial remote sensing platforms (Olsoy *et al.* 2020). Ultimately, the community patterns that emerge from analyzing features across scales could deepen our understanding of plant-herbivore interactions and identify molecular, organismal, and landscape targets for management in changing landscapes.

Overview of Latent Dirichlet Allocation

The overall objective of LDA is to identify latent communities of co-occurring features in data. Communities are latent because they are not directly observed in relative abundance data; instead, communities represent a hidden structure that can be uncovered with statistical modeling. For example, consider a book as a sampling unit, filled with words as features. The co-occurrence of particular words (*e.g.*, "spaceship," "alien," "planet") indicate that book is likely to represent a particular topic, or community (*e.g.*, science fiction). LDA assigns both individual words and individual books to latent communities. LDA factorizes relative abundance data into two matrices, one representing membership of communities in sampling units and the other representing feature membership in communities (Box 1). Input relative abundance data can either represent binary (zero or one) or multinomial (count) data (Valle *et al.* 2018). The number of communities in LDA can either be set in advance or estimated from the data (Albuquerque *et al.* 2019) LDA output includes probabilities of community membership for each sampling unit and feature. As a

generative model (see Box 2), LDA can account for missing data, predict relative abundance at new sites, and represent uncertainty in community membership. For further technical details on LDA, we refer readers to recent reviews by Sankaran and Holmes (2017) and Valle et al. (2014).

Box 1: Modeling Framework

Latent Dirichlet Allocation (LDA) is a statistical model for identifying latent (unobserved) communities. Input data structure for LDA consists of an abundance matrix organized with sampling units (m) as rows and features (n) as columns. Sampling units contain measurements of features. Features measured in sampling units could include species abundance, chemical concentration, or sets of reads from DNA sequencing. Count data on abundance of features in each sample unit is modeled using a multinomial version of LDA:

Abundance of feature *n* in sample unit *m*	$Y_{m,n} \sim Multinomial(\phi_{z[m,n]}, z_{m,n})$;
Abundance of community *z* in sample unit *m*	$z_{m,n} \sim Multinomial(\theta_m, S_{max})$;
Membership of features in communities	$\phi_k \sim Dirichlet(\beta)$;
Membership of communities in sample units	$\theta_m \sim Dirichlet(\gamma)$

where ($Y_{m,n}$) represents the observed abundance of *n*-th feature in *m*-th sample unit. Each entry in the data matrix is assigned to a community type, which is estimated as a latent variable $z_{m,n}$ and depends on the distribution of communities across sample unit *m* (θ_m), and the maximum possible abundance of features in a site (S_{max}). The Dirichlet distribution is a probability distribution for proportional data (Douma and Weedon 2019) that enables mixed membership. The θ_m and ϕ_k parameter matrices reveal the probability that sample units and features, respectively, belong to *k* communities. The hyperparameters, β and γ, represent the degree of mixed membership in the Dirichlet distribution and are often specified to initiate the model (Appendix S2). An alternate parameterization of LDA for binary data assumes that observations are drawn from a binomial distribution but is otherwise similar to the multinomial model (Valle *et al.* 2018b). LDA can be fit with frequentist maximum likelihood estimation or with Bayesian approaches, such as Gibbs sampling (Hornik and Grün 2011). See Supporting Information for more details, including R scripts with examples of LDA fit to multiple types of data.

Applying Latent Dirichlet Allocation Across Data Sets

We analyzed each of our datasets with LDA models in RStudio (v. 3.4.4) to detect communities within sampling units. Our models applied a Bayesian framework from the 'Rlda' package (Albuquerque *et al.* 2019). We used a binomial version of the LDA (Valle *et al.* 2018) to detect communities from occurrence data on metabolites and spectral reflectance, and a multinomial parametrization (Blei *et al.* 2003; Valle *et al.* 2014) for the analyses of count data on microbial taxa and leaf area index. We provide detailed methods for each case study in the Supplementary material (Appendix S1).

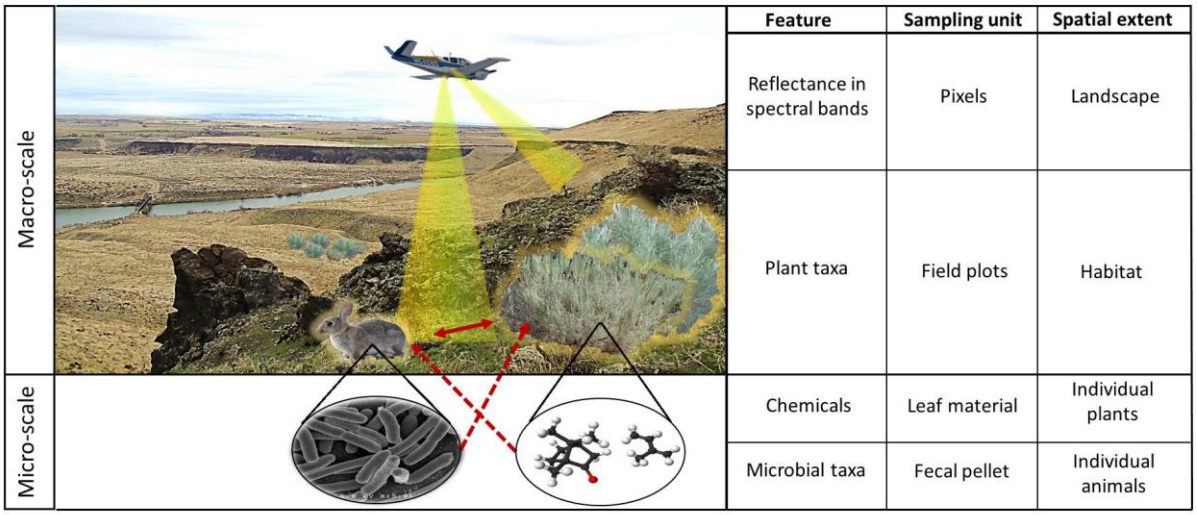


Figure 1: Illustration of how communities are measured in sampling units that span micro- and macro-scales. In the sagebrush steppe ecosystem, these communities are linked across scales. Microbial taxa in fecal pellets from individual herbivores interact with chemical features in leaf material when herbivores consume individual plants. Metabolite features in leaf material consumed by herbivores are dependent on the abundance of individual plant taxa detected within field plots. The distribution of plant taxa can be detected with spectral bands in pixels of aerial imagery obtained remotely within landscapes.

Case Study 1. Reflectance of Spectral Bands within Pixels at the Landscape Scale.

Our first case study uses LDA to assess patterns in spectral data obtained from remotely sensed images. In this case study, communities represent co-occurring wavelengths of spectral reflectance. Understanding impacts of global change on sagebrush ecosystems will require measurements over areas larger than that provided by field plots alone. We used a binomial version of LDA to detect communities or patterns in reflectance from aerial imagery of a sagebrush steppe landscape (Data available from National Ecological Observatory Network. 2019, Figure 2a).

Using LDA, we were able to detect ecological patterns related to changing composition of plants. We identified two communities of spectral features characteristic of vegetation (Figure 2a). Based on visual interpretation of concurrently collected Red-Green-Blue (RGB) imagery, the first community (Community 1) represents juniper trees (*Juniperus* spp.) while the second community is associated with low-growing shrubs (Figure 2b). Juniper range expansion threatens wildlife species (Severson *et al.* 2017). The patchiness of Community 1 suggests fine-scale variation in juniper cover during the early stages of woody encroachment (Figure 2b). The more uniform representation of Community 2 (Figure 2b, right panel) is attributable to a dominant but sparse canopy of shrubs documented in ground observations. Our results demonstrate how high-resolution hyperspectral data can detect and map juniper encroachment in sagebrush steppe. Ultimately, patterns of remotely-sensed, spectral features could be used to monitor ecological change in landscapes where herbivores forage (Frye *et al.* 2013; Ulappa *et al.* 2014).

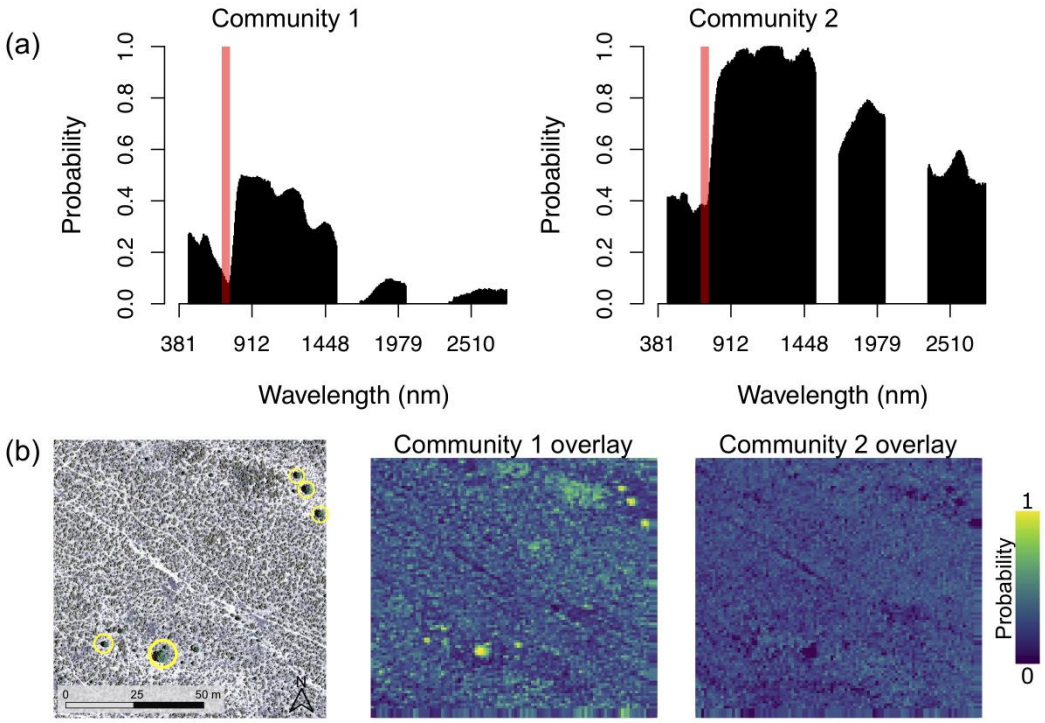


Figure 2: LDA applied to a subset of hyperspectral (1 m^2 resolution) orthomosaic from a sagebrush steppe ecosystem (Onaqui, Utah, USA). (a) The probability of each wavelength of reflected light (nm) belonging to two communities. The rapid change in reflectance between 690 nm and 750 nm (the "red edge") is representative of changes in plant photosynthetic activity. (b) Red, Green, Blue (RGB) image of the area with encroaching juniper trees (*Juniperus* spp.) circled in yellow (left image) with community 1 overlay (middle) and community 2 overlay (right) in the same area outlining a high probability that junipers belong to Community 1. Colors approaching yellow in the community overlays indicate higher probability of pixel membership from a particular spectral feature.

Case Study 2. Plants within Field Plots at the Habitat Scale.

Our second case study uses LDA to detect communities of plant species in field plots using measurements of leaf-area-index (LAI). LAI is the relative size of one leaf over a unit of ground surface (Ewert 2004). Measurements of LAI in drylands relate to food availability for herbivores (Olsoy *et al.* 2015). We quantified LAI in field plots within a Wyoming big sagebrush habitat (*Artemisia tridentata ssp. wyomingensis*) (Figure 3).

Results from applying LDA suggest that this habitat type contained six plant species communities (Figure 3). We report on the composition of three of these communities due to their ecological significance. Community 1 and 3 were dominated by the presence of Wyoming big sagebrush and a native bunchgrass Sandberg bluegrass (*Poa secunda*), respectively. Wyoming big sagebrush and Sandberg bluegrass are of particular importance because their presence indicates habitats favorable for herbivores (Beck *et al.* 2009). Community 6 was dominated by cheatgrass (*Bromus tectorum*), an invasive annual (Figure 3a) that indicates degraded ecosystems less suitable for herbivores (Steenvoorden *et al.* 2019).

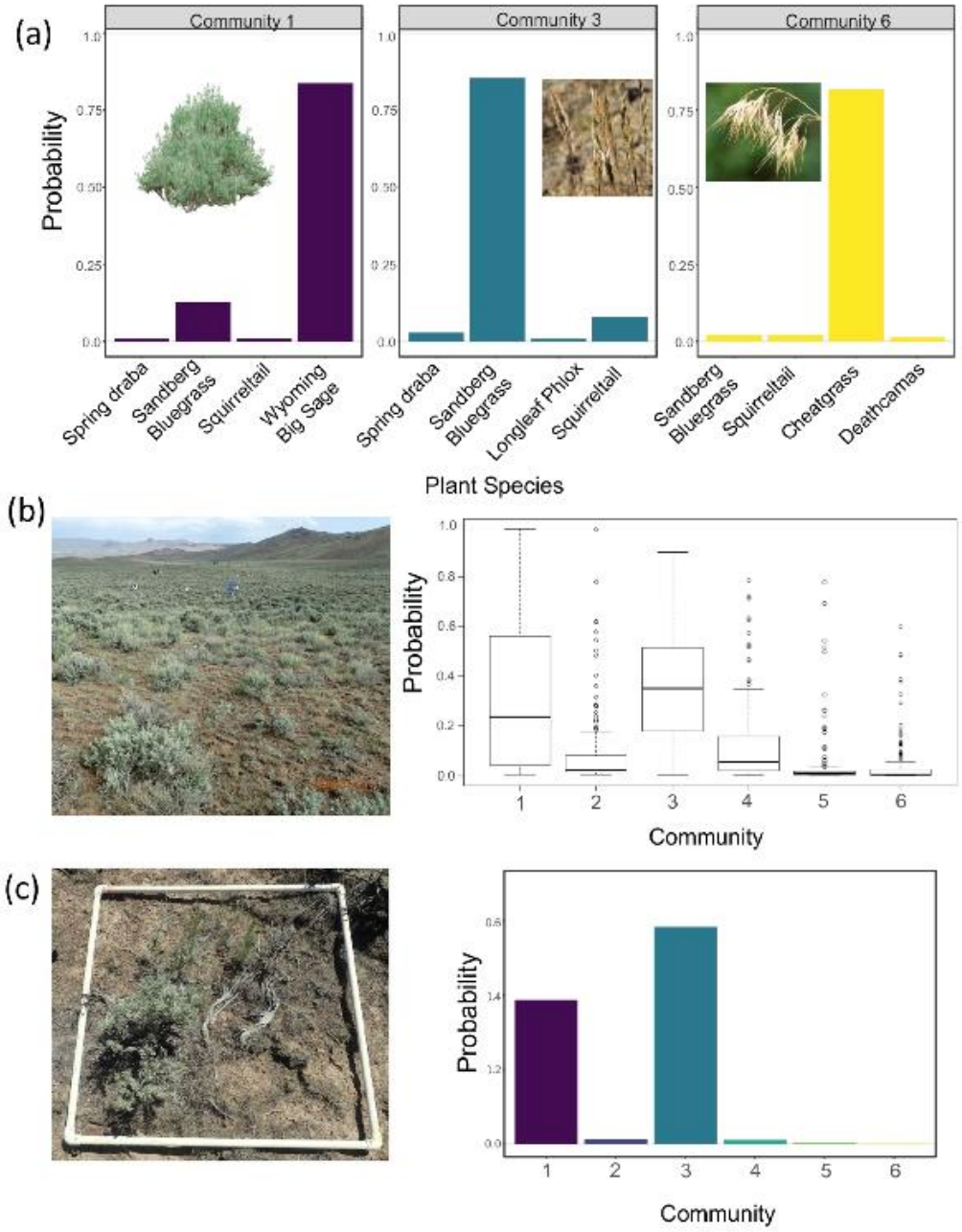


Figure 3. The results of LDA analysis on leaf area index (LAI) in a Wyoming big sagebrush habitat. (a) The probability of plant species occurring within three communities with an image of the dominant species, *Artemisia tridentata ssp. wyomingensis*, in inset. (b) A landscape level photo (left) and the probability of presence of the six most common communities within the habitat sampling units (right). (c) A representative photo of a single $1m^2$ field plot sampling unit (left) and the probability of presence of each community within a single plot (right).

Our results show that communities in this habitat are dominated by Wyoming big sagebrush and Sandberg bluegrass, with low probability of the invasive cheatgrass community (Figure 3b). These results are visible at the level of a single sampling unit ($1m^2$, Figure 3c). Our leaf-level analysis could be used to quantify fine-scale suitability for particular

wildlife species. For example, plots monitored after fires with high probability of Wyoming big sagebrush and low probability of cheatgrass might indicate successful post-fire restoration (Baker 2006), including the regeneration of suitable forage for herbivores (Beck *et al.* 2012).

Case Study 3. Metabolites within Leaves at the Plant Scale.

Our third case study uses LDA to detect patterns in metabolite features, specifically volatile monoterpenes, between two sagebrush taxa. While several herbivores rely on sagebrush as forage year-round, the volatile monoterpene features of this plant influence selection by herbivores at the species, patch, and plant scale (Frye *et al.* 2013). Although there are known concentration-dependent consequences of individual monoterpenes, the unique mixtures of metabolites in plants may better explain intake by herbivores (Nobler *et al.* 2019). Moreover, foraging herbivores consume mixtures of metabolites, not individual metabolites. Approaches that focus on the presence or concentration of a specific metabolite likely miss differences in the relative ratios of compounds that better determine diet selection by herbivores and predict interactions with the microbial features (*e.g.*, case study 4 below) in herbivore guts. In this case study, communities represent "chemical bouquets," or groups of secondary metabolites.

We found that LDA detected communities of monoterpenes that were relevant to herbivore diet selection in two different sagebrush taxa (Figure 4). Specifically, there were three communities that contained unique individual monoterpenes known to predict foraging by herbivores. While the identity of several monoterpenes quantified are unknown (Unk indicate unknown compounds), the suite of monoterpenes in Community 1, 3, and 4 were present across many of the plant samples (Figure 4b). At the individual sampling unit (plant) level, three-tip sagebrush (*Artemisia tripartita*) had a high probability of Community 4, whereas Wyoming big sagebrush was dominated by Communities 1 and 3 (Figure 4c). Concentrations of Unk 21.0 (dominant in Community 4) and Unk 21.5 (dominant in Community 3) have previously been found to predict diet selection by free-ranging sage-grouse (Fremgen-Tarantino *et al.* 2020) and β-pinene (dominant in Community 1) was avoided by captive mountain cottontails (*Sylvilagus nuttallii*) (Nobler *et al.* 2019). Our results demonstrate how LDA can reveal communities of metabolite features that predict foraging decisions by herbivores. A potential application of LDA could be to improve post-fire restoration by reseeding with plants that have similar chemical community profiles to plants consumed and preferred by threatened herbivores.

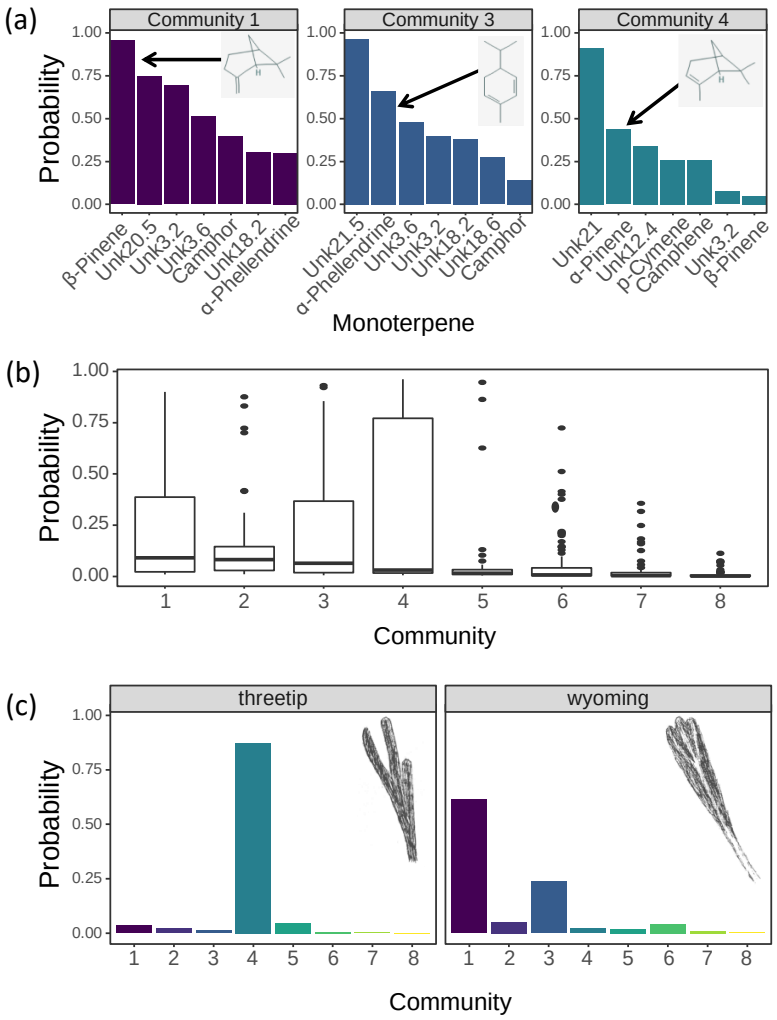


Figure 4. Results of LDA analysis of monoterpenes from leaves of sagebrush plants consumed by herbivores. (a) Probability of monoterpenes occurring within three metabolite communities with an image of the molecular structure of the dominate known monoterpene in inset. (b) Probability of the eight most common metabolite communities across all sagebrush samples. (c) Probability of the eight metabolite communities occurring within an individual three-tip (*Artemisia tripartita*) and a Wyoming big sagebrush (*Artemisia tridentata ssp. wyomingensis*) sampling unit with an image of the leaf morphology of each species in inset.

Case Study 4. Microbial Taxa within Fecal Pellets at the Herbivore Scale.

Chemical communities in herbivore forage, including plants in the wild and artificial pellets in captivity, can modify microbial species composition within animal guts (Kohl *et al.* 2014; Sandifer *et al.* 2015; Mohajeri *et al.* 2018). Our fourth case study uses LDA to identify patterns in microbial taxonomic features detected in fecal pellets of pygmy rabbits over time. For this case study, communities represent groups of co-occurring microbial taxa. Specifically, we analyzed how the taxa of the fecal microbiome from this obligate sagebrush herbivore would change as they were transitioned from a natural diet containing Wyoming big sagebrush to a captive diet, containing commercial rabbit food, over a seven-day period. Fecal samples from the rabbits on day 1 (sagebrush diet) and day 10 (captive diet) were collected and analyzed using shotgun metagenomics. We used LDA to identify communities of bacteria at the genus level (Figure 5). The anaerobes, *Clostridium* and *Bacteroides*, were common features of these bacterial communities (Figure 5a). Communities 3 and 8 show the highest probability of being found within all fecal samples (Figure 5b). Community 3 was dominated by *Bacteroides* and had a higher probability of being present when the rabbits were on a natural diet, whereas Community 8, which was dominated by *Clostridium* species, was more prevalent after a week

of transitioning to a captive diet (Figure 5c). Some *Clostridium* species are associated with enteritis (inflammation of the small intestine) and increased mortality in wild and captive animals (Paul and Friend 2019) whereas other *Clostridium* species may improve animal health (Liu *et al.* 2019). These preliminary results suggest that LDA can be used to monitor changes in bacterial communities associated with dietary shifts, and potential health, in sagebrush-dependent herbivores. Because microbial function is largely driven by communities, rather than individual species, community-level analyses (*e.g.*, LDA) are crucial for identifying physiologically-relevant changes in herbivore metagenomes.

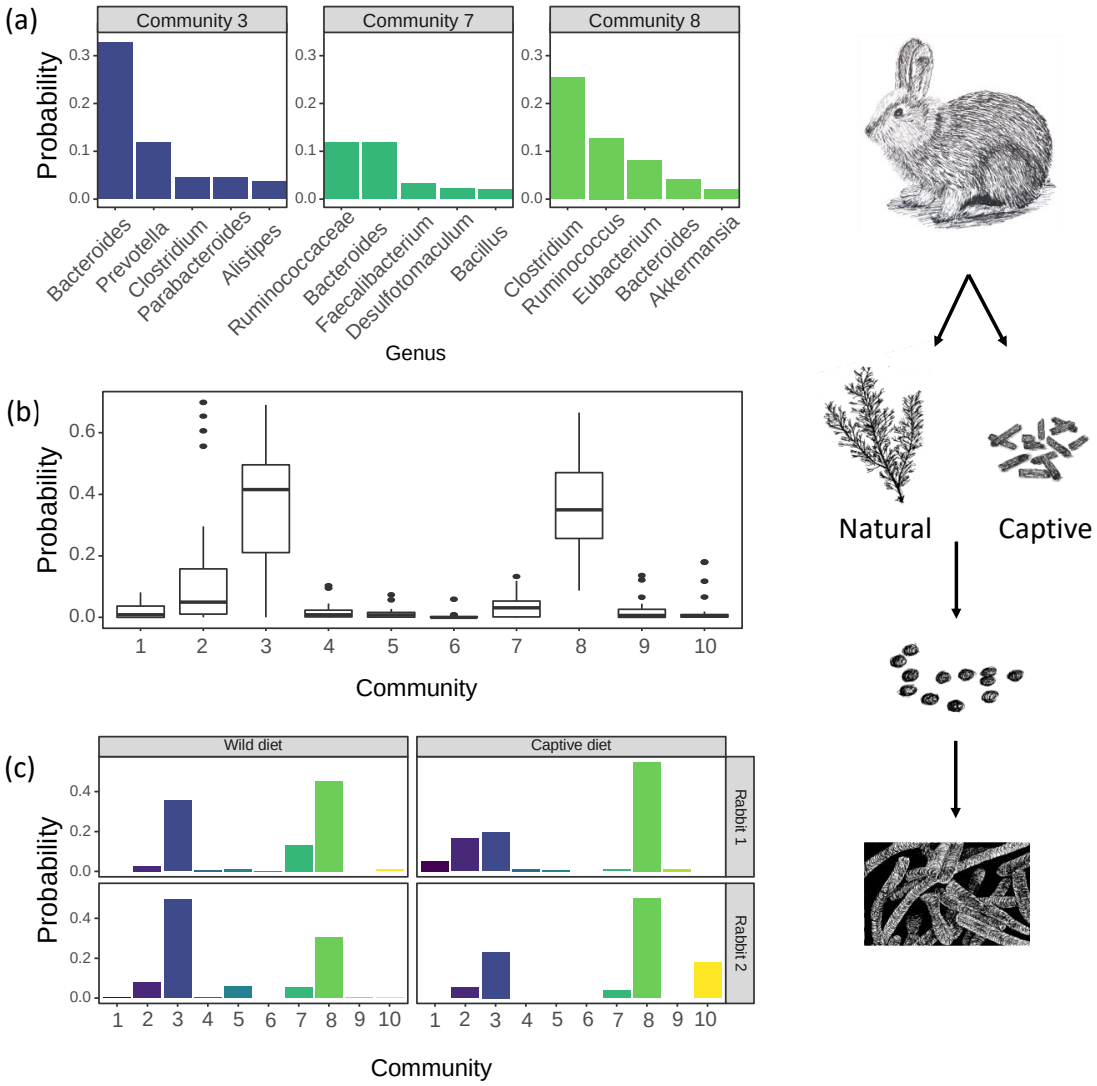


Figure 5. LDA analysis using genus level taxonomy counts from metagenomics of fecal samples collected from pygmy rabbits (*Brachylagus idahoensis*, shown top left). (a) Probability of microbial features within the three most prevalent communities detected in fecal samples, each dominated by different microbial taxa. (b) Probability of the ten identified microbial communities within fecal samples from the pygmy rabbit (n=22). (c) Probability of the ten microbial communities in fecal samples from the pygmy rabbit sampling units consuming a natural diet (primarily Wyoming big sagebrush (*Artemisia tridentata ssp. wyomingensis*) and after seven days on an artificial pellet diet in captivity.

Discussion

As biodiversity science grows to encompass scales from molecular to continental, the need for integrative approaches has increased as well. We have demonstrated the potential for community detection to unite patterns of biodiversity across disciplines. Latent Dirichlet Allocation, a topic model that can represent mixed membership of features, enabled us to quantify communities across molecular, organismal, and landscape scales. Our results have potential relevance for conservation of threatened herbivores in the sagebrush steppe ecosystem, including the pygmy rabbit. At the landscape scale, LDA detected juniper encroachment, a driver of habitat degradation in sagebrush steppe, from aerial remote sensing data. At the plant scale, LDA enabled discrimination between plant species assemblages, with relevance for habitat structure, including the availability of quality forage for herbivores and the presence of invasive species. At the molecular scale, LDA identified mixtures of secondary metabolites that can differentiate plant species and predict diet selection by herbivores. At the microbial scale, LDA quantified shifts in bacterial communities that are predictive of disease and survival, and respond to diet transitions of herbivores. Co-analyzing datasets with LDA improved comprehension of biodiversity across scales for members of our multidisciplinary research team, leading us to develop a more holistic view of plant-herbivore ecology. Common models for disparate datasets, including LDA, will enable collaborative studies that can better inform cross-scale strategies for conservation.

One realization that emerged from co-analyzing our data is the importance of herbivore gut microbiomes for uniting scales. We argue that studying gut microbiomes has great potential to develop a more complete understanding of herbivore ecology, particularly if multiple scales are incorporated into analyses. Herbivores, such as the pygmy rabbits in our study, make foraging decisions at individual metabolite, leaf, plant, and landscape scales (Ulappa *et al.* 2014; Nobler *et al.* 2019). In turn, foraging herbivores can influence patterns of habitat structure and plant species composition (Eldridge *et al.* 2016). Over long periods of time, we would expect that gut microbes mediate feedback loops between plants and herbivores, with ecological and evolutionary implications (Ley *et al.* 2008; Kohl and Dearing 2016). In a practical sense, the gut microbiome links these disparate scales and represents the net sum of forage availability and quality across landscapes (Figure 2), habitat patches (Figure 3) and within plants (Figure 4). Herbivore foraging has wide-ranging consequences for above-ground (Frye *et al.* 2013; Ulappa *et al.* 2014; Fremgen-Tarantino *et al.* 2020) and below-ground (Chomel *et al.* 2016) ecological processes, therefore a more holistic understanding of co-occurring plant, metabolite, and microbial communities in the guts of herbivores is needed. Topic models, such as LDA, present an opportunity to describe microbial community structure (Chen *et al.* 2012). The development of analytical tools that integrate hierarchies of scale and complex network structure will further enable researchers to uncover how microbial communities might interact with communities at other scales, from molecular to landscape.

Given the importance of the gut microbiome, we envision designing future studies to collect data from multiple biological units at the same time and place with a focus around fecal collections. Data collection focused around herbivore fecal pellets could involve sampling feces from herbivores for metagenomic and metabolite analyses while simultaneously collecting leaf tissue from plants browsed by herbivores for metabolite content (parent and detoxification products), and mapping the GPS location where pellets and plant samples are collected. Subsequently, research teams could assess how communities of microbes, plant-derived metabolites, and plant species detected in feces are influenced by variation in plant species availability at the scale of foraging plots. Remote sensing data, such as hyperspectral aerial images, could then be applied to detect temporal and spatial variation in the composition of plant species and foliar chemistry across the landscape (Fine *et al.* 2021). This type of data collection will require extensive interdisciplinary coordination, but will lead to a more connected understanding of coupled biodiversity among scales. Long-term ecological research sites, such as the NEON network, provide a valuable starting point for this type of study where collection and analyses of herbivore metagenomics and metabolites from plants could add substantial value to existing data on plant diversity and soil microbial communities. In the context of planning field studies, LDA could be applied as a generative model to simulate data and estimate appropriate sample sizes for statistical inference (Box 2).

LDA has distinct advantages over other clustering techniques. A key strength of LDA is its probabilistic nature, enabling detection of novel communities. For example, when analyzing metabolomic data sets from tandem mass spectrometry analysis, LDA identified relevant substructures from co-occurrence of mass fragments and neutral losses in 70% of spectra analyzed, in contrast to other clustering techniques that only found hits for 25 and 6% of spectra respectively (van der Hooft 2016). In some cases, LDA may also improve classification accuracy. For example, LDA-based methods outperformed simple harmonization methods based on semantic affinity scores for identifying latent land cover communities from different source maps (Li *et al.* 2020). More broadly, continuous representations of

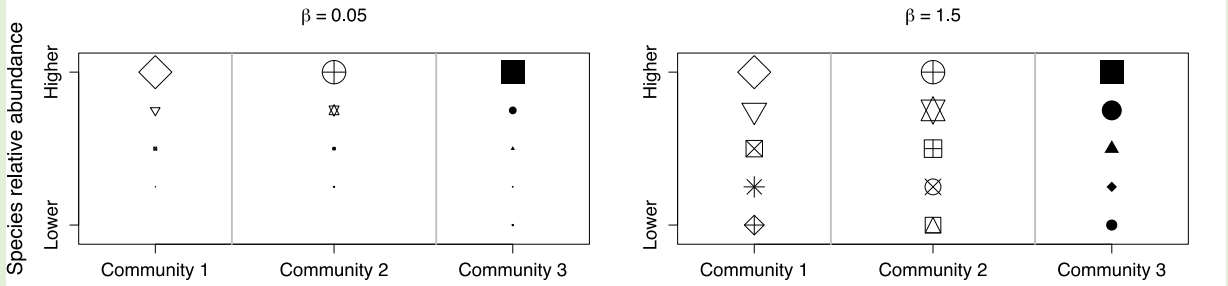
community membership from LDA enable explanatory analyses that would not be possible with clustering methods that assume discrete membership (Knott *et al.* 2020). In addition, LDA is a generative model that can simulate data to improve links between ecological theory and statistical analyses (Box 2). Altogether, probabilistic topic models for community detection are poised to generate novel insights from biodiversity data.

Box 2: Latent Dirichlet Allocation as a Generative Model for Biodiversity

LDA belongs to a broad class of models known as *generative models* that define joint probabilities for latent variables and observed features (Bernardo *et al.* 2007), including a data generating mechanism for observed data based on probability distributions. In contrast, commonly used models for multivariate data in ecology (*e.g.*, ordination; Legendre and Legendre 2012) tend to be non-generative, describing patterns in community data without a probabilistic explanation. A key advantage of generative models is the ability to simulate data that is consistent with observed data. Data simulation from generative models is increasingly considered best practice for statistical analyses (Gabry *et al.* 2019). Comparing observed data to simulations improves overall understanding of data, can identify potential pathologies in statistical models, and assists the design of more efficient sampling schemes.

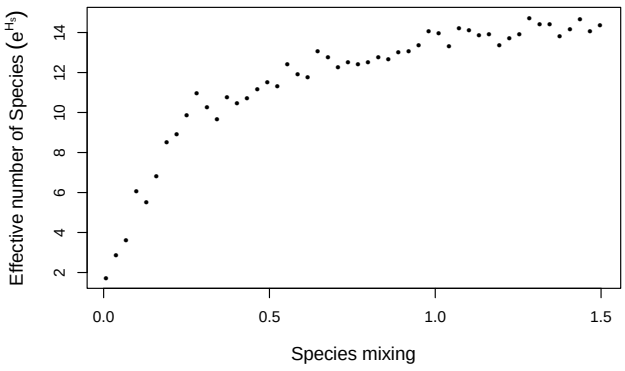
Generative models can also provide a link to ecological theory (Harris *et al.* 2017). In community ecology, long-standing debates on biodiversity metrics for alpha, beta, and gamma diversity have relied upon simulations for understanding these metrics (Legendre *et al.* 2005; Baselga 2010; Veech and Crist 2010). However, simulation models used to explore general properties of biodiversity metrics are often not the same models used to analyze observed data; LDA presents an opportunity to better integrate simulation experiments and statistical models.

To demonstrate the potential for LDA to simulate biodiversity data, we generated 50 fake datasets from the multinomial model described in Box 1. For each dataset, we varied the hyperparameter for membership of species (*i.e.*, features) in latent communities (β in Box 1) from $\beta = 0.01$ to $\beta = 1.5$. A heuristic explanation for this hyperparameter is the degree of species mixing within communities. A lower β value corresponds to a minimal mixing of species in communities and results in a few high-abundance species. In contrast, with increasing β, there is a high degree of species mixing resulting in higher probability of a more uniform distribution of species abundances. Note, the datasets were generated with the site mixing hyperparameter (γ in Box 1) set to zero, resulting in no mixing across communities (high species turnover). In practice, however, the observed patterns of species abundances within and across sites jointly depend on both β and γ hyperparameters.



The figure above shows two simulated datasets on species relative abundance, generated by LDA. The left panel shows three communities from a simulated dataset with low species mixing in and the right panel shows three communities from a simulated dataset with high mixing. Each icon represents one species, with the relative size of each icon indicative of relative abundance of that species within a community.

After simulating datasets using LDA, we then calculated alpha diversity, the variation in species composition within sites, for each of the simulated communities using the Expected Number of Species ($e^{H_{Shannon}}$). Results from this simulation experiment demonstrate a strong relationship between a commonly-used metric for alpha diversity and LDA. These results demonstrate how using LDA as a generative model could provide a way to better understand fundamental concepts in community ecology by linking statistical models for observed data with simulation experiments.



The hyperparameter for species mixing in LDA (β in Box 1) provides a generative model for biodiversity metrics, including alpha diversity. Each dot represents one simulation, with a different hyperparameter for species mixing.

Continued Advances in Community Detection Across Scales

While our case studies of community membership represent separate analyses, topic modeling is well-poised to address long-standing questions of whether different taxonomic units co-occur in space and time (Heino 2010). Correlated topic models represent covariance between communities (Blei and Lafferty 2007), using mathematical relationships that are similar to existing models for co-occurrence between species in joint species distribution models (Pollock *et al.* 2014). Correlated topic models could quantify whether species from different trophic levels co-occur in space (often referred to as spatial concordance or cross-taxon congruence; Pearson and Carroll 1999; Su *et al.* 2004). Analyses of cross-taxon congruence typically involves a two-step process, first to identify latent communities and second to analyze correlations between them (Heino 2010). Correlated topic models present an opportunity to combine both of these steps into a single statistical model. The flexibility of probabilistic models, such as LDA, could prove invaluable for extending questions of cross-taxon congruence beyond species to secondary metabolites, genes, landscape features, and other levels of biological organization.

Biodiversity data commonly includes features and communities that change over time and in response to environmental covariates. Newly developing topic modeling approaches could improve our capacity for inference on drivers of community membership. Dynamic topic models are currently used in text mining to account for changing community membership (Blei and Lafferty 2006), while dynamic mixture models enable realized proportions of communities to change over time (Wei *et al.* 2007). In ecology, LDA has recently been applied to interpret how gradual changes in rodent communities over a 40-year period were related to environmental drivers (Christensen *et al.* 2018). As a statistical approach conceptually related to regression models for proportional data (Douma and Weedon 2019), LDA can provide insights on community dynamics across time and space.

Conclusions

Coordinated studies of community structure across scales will enable researchers to address fundamental questions in ecology and evolution. One such question relates to the long-standing debate over whether biological features, from genes to species assemblages, are organized by neutral processes or deterministic ecological and evolutionary forces (Kreitman 1996; Lynch 2007; Lowe and McPeek 2014). For example, convergent communities of microbes in the soil and guts of herbivores exposed to similar plant metabolite communities across broad biogeographical scales would provide powerful evidence for the role of non-neutral processes. Alternately, random associations between overlain communities could suggest neutral theory as an explanation for observed assemblages. Common models for community structure will provide detailed and cohesive insight into the complex interactions among plants, animals, and microbes co-occurring across landscapes. Altogether, we anticipate that interdisciplinary collaboration, facilitated by the common modeling language of LDA, will have payoffs for biodiversity studies that must address complex problems across scales.

Acknowledgements

We thank J. Connelly, D. D. Musil, L. Cross, L.A. Shipley and the GUTT seminar working group in addition to financial support from a NASA grant 80NSCCC17K0738, a Idaho State Board of Education grant IGEM19-002, a Semiconductor Research Corporation grant SRC 2018-SB-2842, Pittman-Robertson 683 funds from the Idaho Department of Fish and Game, Sigma Xi Grants-In-Aid, Bureau of Land Management grant #L09AC16253, the USDA Agricultural Research Service and National Science Foundation grants IOS-1258217, DEB-1146194, DEB-1146368, OIA-1826801, OIA-1757324, OIA-1738865 and ECCS-1807809. Thanks to James Hudon for drawings in Figures 4 and 5.

References

Albuquerque PHM, Valle DR do, and Li D. 2019a. Bayesian LDA for mixed-membership clustering analysis: The Rlda package. *Knowl-Based Syst* **163**: 988–95.

Albuquerque PHM, Valle DR do, and Li D. 2019b. Bayesian LDA for mixed-membership clustering analysis: The Rlda package. *Knowl-Based Syst* **163**: 988–95.

Baker WL. 2006. Fire and Restoration of Sagebrush Ecosystems. *Wildl Soc Bull* **34**: 177–85.

Barde, BV, Bainwad, AM. An overview of topic modeling methods and tools. In: IEEE international conference on intelligent computing and control systems (ICICCS), Madurai, India, 15–16 June 2017, pp. 745–750. New York: IEEE.

Baselga A. 2010. Multiplicative partition of true diversity yields independent alpha and beta components; additive partition does not. *Ecology* **91**: 1974–81.

Beck JL, Connelly JW, and Reese KP. 2009. Recovery of Greater Sage-Grouse Habitat Features in Wyoming Big Sagebrush following Prescribed Fire. *Restor Ecol* **17**: 393–403.

Beck JL, Connelly JW, and Wambolt CL. 2012. Consequences of Treating Wyoming Big Sagebrush to Enhance Wildlife Habitats. *Rangel Ecol Manag* **65**: 444–55.

Bennett EM, Cramer W, Begossi A, *et al.* 2015. Linking biodiversity, ecosystem services, and human well-being: three challenges for designing research for sustainability. *Curr Opin Environ Sustain* **14**: 76–85.

Bernardo JM, Bayarri MJ, Berger JO, *et al.* 2007. Generative or discriminative? getting the best of both worlds. *Bayesian Stat* **8**: 3–24.

Blei DM and Lafferty JD. 2006. Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery.

Blei DM and Lafferty JD. 2007. A correlated topic model of science. *Ann Appl Stat* **1**: 17–35.

Blei DM, Ng AY, and Jordan MI. 2003. Latent Dirichlet Allocation. *J Mach Learn Res* **3**: 993–1022.

Burke C, Steinberg P, Rusch D, *et al.* 2011. Bacterial community assembly based on functional genes rather than species. *Proc Natl Acad Sci* **108**: 14288–93.

Chen X, Hu X, Lim TY, *et al.* 2012. Exploiting the functional and taxonomic structure of genomic data by probabilistic topic modeling. *IEEE/ACM Trans Comput Biol Bioinform* **9**: 980–91.

Chen X, Hu X, Shen X, and Rosen G. 2010. Probabilistic topic modeling for genomic data interpretation. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE.

Chen X, Hu X, Shen X, Rosen G (2010) Probabilistic topic modeling for genomic data interpretation. IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE, Piscataway, NJ), pp 149–152.

Chomel M, Guittonny-Larchevêque M, Fernandez C, *et al.* 2016. Plant secondary metabolites: a key driver of litter decomposition and soil nutrient cycling. *J Ecol* **104**: 1527–41.

Christensen EM, Harris DJ, and Ernest SKM. 2018. Long-term community change through multiple rapid transitions in a desert rodent community. *Ecology* **99**: 1523–9.

Clements FE. 1936. Nature and Structure of the Climax. *J Ecol* **24**: 252.

Cross, T. B. et al. 2018. The genetic network of greater sage-grouse: Range-wide identification of keystone hubs of connectivity. - Ecology and evolution 8: 5394–5412.

Der Hooft JJJ van, Wandy J, Barrett MP, *et al.* 2016. Topic modeling for untargeted substructure exploration in metabolomics. *Proc Natl Acad Sci* **113**: 13738–43.

Douma JC and Weedon JT. 2019. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods Ecol Evol* **10**: 1412–30.

Eldridge DJ, Poore AGB, Ruiz-Colmenero M, *et al.* 2016. Ecosystem structure, function, and composition in rangelands are negatively affected by livestock grazing. *Ecol Appl* **26**: 1273–83.

Ewert F. 2004. Modelling Plant Responses to Elevated CO2: How Important is Leaf Area Index? *Ann Bot* **93**: 619–27.

Fauchald P, Park T, Tømmervik H, *et al.* 2017. Arctic greening from warming promotes declines in caribou populations. *Sci Adv* **3**: e1601365.

Fine PVA, Salazar D, Martin RE, *et al.* 2021. Exploring the links between secondary metabolites and leaf spectral reflectance in a diverse genus of Amazonian trees. *Ecosphere* **12**: e03362.

Fremgen-Tarantino MR, Peña JJ, Connelly JW, and Forbey JS. 2020. Winter foraging ecology of Greater Sage-Grouse in a post-fire landscape. *J Arid Environ* **178**: 104154.

Frye GG, Connelly JW, Musil DD, and Forbey JS. 2013. Phytochemistry predicts habitat selection by an avian herbivore at multiple spatial scales. *Ecology* **94**: 308–14.

Gabry J, Simpson D, Vehtari A, *et al.* 2019. Visualization in Bayesian workflow. *J R Stat Soc Ser A Stat Soc* **182**: 389–402.

Gleason HA. 1926. The Individualistic Concept of the Plant Association. *Bull Torrey Bot Club* **53**: 7–26.

Harris K, Parsons TL, Ijaz UZ, *et al.* 2017. Linking Statistical and Ecological Theory: Hubbell's Unified Neutral Theory of Biodiversity as a Hierarchical Dirichlet Process. *Proc IEEE* **105**: 516–29.

Heino J. 2010. Are indicator groups and cross-taxon congruence useful for predicting biodiversity in aquatic ecosystems? *Ecol Indic* **10**: 112–7.

Hornik K and Grün B. 2011. topicmodels: An R package for fitting topic models. *J Stat Softw* **40**: 1–30.

Kennedy RE, Townsend PA, Gross JE, *et al.* 2009. Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects. *Remote Sens Environ* **113**: 1382–96.

Knott, J. A. et al. 2020. Community-level responses to climate change in forests of the eastern United States. - Global Ecology and Biogeography 29: 1299–1314.

Kohl, K. D. et al. 2016. Microbial detoxification in the gut of a specialist avian herbivore, the Greater Sage-Grouse. - FEMS Microbiology Letters 363: fnw144. https://doi.org/10.1093/femsle/fnw144

Kohl KD and Dearing MD. 2016. The Woodrat Gut Microbiota as an Experimental System for Understanding Microbial Metabolism of Dietary Toxins. *Front Microbiol* **7**.

Kohl KD, Weiss RB, Cox J, *et al.* 2014. Gut microbes of mammalian herbivores facilitate intake of plant toxins. *Ecol Lett* **17**: 1238–46.

Kreitman M. 1996. The neutral theory is dead. Long live the neutral theory. *Bioessays* **18**: 678–83.

Legendre P, Borcard D, and Peres-Neto PR. 2005. Analyzing Beta Diversity: Partitioning the Spatial Variation of Community Composition Data. *Ecol Monogr* **75**: 435–50.

Legendre P and Legendre L. 2012. Numerical Ecology. Elsevier.

Ley RE, Hamady M, Lozupone C, *et al.* 2008. Evolution of Mammals and Their Gut Microbes. *Science* **320**: 1647–51.

Li Z, White JC, Wulder MA, *et al.* 2020. Land cover harmonization using Latent Dirichlet Allocation. *Int J Geogr Inf Sci* **0**: 1–27.

Liu L, Zeng D, Yang M, *et al.* 2019. Probiotic Clostridium butyricum Improves the Growth Performance, Immune Function, and Gut Microbiota of Weaning Rex Rabbits. *Probiotics Antimicrob Proteins* **11**: 1278–92.

Lortie CJ, Brooker RW, Choler P, *et al.* 2004. Rethinking plant community theory. *Oikos* **107**: 433–8.

Lowe WH and McPeek MA. 2014. Is dispersal neutral? *Trends Ecol Evol* **29**: 444–50.

Lynch M. 2007. The evolution of genetic networks by non-adaptive processes. *Nat Rev Genet* **8**: 803–13.

McCune B, Grace JB, and Urban DL. 2002. Analysis of ecological communities. MjM software design Gleneden Beach, OR.

Mohajeri MH, Brummer RJM, Rastall RA, *et al.* 2018. The role of the microbiome for human health: from basic science to clinical applications. *Eur J Nutr* **57**: 1–14.

Mosher BA, Bernard RF, Lorch JM, *et al.* 2020. Successful molecular detection studies require clear communication among diverse research partners. *Front Ecol Environ* **18**: 43–51.

National Ecological Observatory Network. 2019. Data Product DP3.30006.001, Spectrometer orthorectified surface directional reflectance - mosaic. Provisional data downloaded from http://data.neonscience.org February 2019.

Nemergut DR, Schmidt SK, Fukami T, *et al.* 2013. Patterns and processes of microbial community assembly. *Microbiol Mol Biol Rev* **77**: 342–56.

Nobler JD, Camp MJ, Crowell MM, *et al.* 2019. Preferences of Specialist and Generalist Mammalian Herbivores for Mixtures Versus Individual Plant Secondary Metabolites. *J Chem Ecol* **45**: 74–85.

Oh KP, Aldridge CL, Forbey JS, *et al.* 2019. Conservation Genomics in the Sagebrush Sea: Population Divergence, Demographic History, and Local Adaptation in Sage-Grouse (Centrocercus spp.). *Genome Biol Evol* **11**: 2023–34.

Olsoy PJ, Forbey JS, Rachlow JL, *et al.* 2015. Fearscapes: Mapping Functional Properties of Cover for Prey with Terrestrial LiDAR. *BioScience* **65**: 74–80.

Olsoy, P. J. et al. 2020. Mapping foodscapes and sagebrush morphotypes with unmanned aerial systems for multiple herbivores. - Landscape Ecology: 1–16.

Paul GC and Friend DG. 2019. Clostridial Enterotoxemia and Coccidiosis in Weanling Cottontail Rabbits (Sylvilagus audubonii, Sylvilagus floridanus, Sylvilagus nuttallii) from Colorado, USA. *J Wildl Dis* **55**: 189–95.

Pearson DL and Carroll SS. 1999. The influence of spatial scale on cross-taxon congruence patterns and prediction accuracy of species richness. *J Biogeogr* **26**: 1079–90.

Pollock LJ, Tingley R, Morris WK, *et al.* 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods Ecol Evol* **5**: 397–406.

Pritchard JK, Stephens M, and Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–59.

Räsänen, A. et al. 2016. The role of landscape, topography, and geodiversity in explaining vascular plant species richness in a fragmented landscape. **21**, 53–70.

Requena-Mullor, J. M. et al. 2019. Integrating anthropogenic factors into regional-scale species distribution models—A novel application in the imperiled sagebrush biome. - Global Change Biology 25: 3844–3858.

Sandifer PA, Sutton-Grier AE, and Ward BP. 2015. Exploring connections among nature, biodiversity, ecosystem services, and human health and well-being: Opportunities to enhance health and biodiversity conservation. *Ecosyst Serv* **12**: 1–15.

Sankaran, K. and Holmes, S. P. 2019. Latent variable modeling for the microbiome. - Biostatistics 20: 599–614.

Severson JP, Hagen CA, Maestas JD, *et al.* 2017. Effects of conifer expansion on greater sage-grouse nesting habitat selection. *J Wildl Manag* **81**: 86–95.

Simpson EH. 1949. Measurement of Diversity. *Nature* **163**: 688–688.

Steenvoorden J, Meddens AJH, Martinez AJ, *et al.* 2019. The potential importance of unburned islands as refugia for the persistence of wildlife species in fire-prone ecosystems. *Ecol Evol* **9**: 8800–12.

Su JC, Debinski DM, Jakubauskas ME, and Kindscher K. 2004. Beyond species richness: Community similarity as a measure of cross-taxon congruence for coarse-filter conservation. *Conserv Biol* **18**: 167–73.

Trevelline BK, Fontaine SS, Hartup BK, and Kohl KD. 2019. Conservation biology needs a microbial renaissance: a call for the consideration of host-associated microbiota in wildlife management practices. *Proc R Soc B Biol Sci* **286**: 20182448.

Ulappa AC, Kelsey RG, Frye GG, *et al.* 2014. Plant protein and secondary metabolites influence diet selection in a mammalian specialist herbivore. *J Mammal* **95**: 834–42.

Valle D, Albuquerque P, Zhao Q, *et al.* 2018a. Extending the Latent Dirichlet Allocation model to presence/absence data: A case study on North American breeding birds and biogeographical shifts expected from climate change. *Glob Change Biol* **24**: 5560–72.

Valle D, Albuquerque P, Zhao Q, *et al.* 2018b. Extending the Latent Dirichlet Allocation model to presence/absence data: A case study on North American breeding birds and biogeographical shifts expected from climate change. *Glob Change Biol* **24**: 5560–72.

Valle D, Baiser B, Woodall CW, and Chazdon R. 2014. Decomposing biodiversity data using the Latent Dirichlet Allocation model, a probabilistic multivariate statistical method. *Ecol Lett* **17**: 1591–601.

Veech JA and Crist TO. 2010. Diversity partitioning without statistical independence of alpha and beta. *Ecology* **91**: 1964–9.

Vellend M. 2010. Conceptual Synthesis in Community Ecology. *Q Rev Biol* **85**: 183–206.

Wei, X. et al. 2007. Dynamic Mixture Models for Multiple Time-Series. - Ijcai 7: 2909–2914.

White SM, Flinders JT, and Welch BL. 1982. Preference of Pygmy Rabbits (Brachylagus idahoensis) for Various Populations of Big Sagebrush (Artemisia tridentata). *J Range Manag* **35**: 724–6.

Zaiats, Andrii et al. (2021), Unifying community detection across scales from genomes to landscapes., Dryad, Dataset, https://doi.org/10.5061/dryad.8w9ghx3mf