Scaling Laws for Function Diversity and Specialization Across Socioeconomic and Biological Complex Systems

Vicky Chuqiao Yang[a,b], James Holehouse[c], Hyejin Youn[c,d,e], José Ignacio Arroyo[c], Sidney Redner[c], Geoffrey B. West[c, f], and Christopher P. Kempes[c]

[a]MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge MA, USA
[b]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge MA, USA
[c]Santa Fe Institute, Santa Fe, NM, USA
[d]Graduate School of Business, Seoul National University, Seoul, South Korea
[e]Northwestern Institute on Complex Systems, Evanston, IL, USA
[f]London School of Economics and Political Science, London, United Kingdom

Abstract

Function diversity, the range of tasks individuals perform, and specialization, the distribution of function abundances, are fundamental to complex adaptive systems. In the absence of overarching principles, these properties have appeared domain-specific. Here, we introduce an empirical framework and a mathematical model for the diversification and specialization of functions across disparate systems, including bacteria, federal agencies, universities, corporations, and cities. We find that the number of functions grows sublinearly with system size, with exponents from 0.35 to 0.57, consistent with Heaps' Law. In contrast, cities exhibit logarithmic scaling. To explain these empirical findings, we generalize the Yule-Simon model by introducing two key parameters: a diversification parameter that characterizes how existing functions inhibit the creation of new ones, and a specialization parameter that describes how a function's attractiveness depends on its abundance. Our model enables cross-system comparisons, from microorganisms to metropolitan areas. The analysis suggests that what drives the creation of new functions depends on the system's goals and structure: federal agencies tend to ensure comprehensive coverage of necessary functions; cities tend to slow the creation of new occupations as existing ones expand; and cells occupy an intermediate position. Once functions are introduced, their growth follows a remarkably universal pattern across all systems.

Significance Statement

Diversification and specialization are central to complex adaptive systems, yet overarching principles across domains remain elusive. We introduce a general theory that unifies diversity and specialization across disparate systems, including microbes, federal agencies, companies, universities, and cities, characterized by two key parameters. We show from extensive data that function diversity scales with system size as a sublinear power law—resembling Heaps' law—in all but cities, where it is logarithmic. Our theory explains both behaviors and suggests that function creation depends on system goals and structure: federal agencies tend to ensure functional coverage; cities slow new function growth as old ones expand, and cells occupy an intermediate position. Once functions are introduced, their growth follows a remarkably universal pattern across all systems.

Introduction

Fundamental to complex adaptive systems—encompassing biological organisms, human organizations, and urban areas—is the range of functions (diversity) and their relative abundance (specialization). Diversity, the number of unique sub-components within a system, indicates a broad range of abilities and capabilities for organisms or organizations to tackle complex challenges, hedge against risks, and thus adapt to complex environments (see Table 1 for some examples)[1–6]. As these systems mature, develop, and expand, their function diversity grows, reflecting an increase in adaptability and resilience [1–3, 7, 8]. Remarkably consistent empirical regularities are observed in both biological and socio-economic systems, despite their distinct environmental conditions and evolutionary trajectories. Examples include the species-area curve—the empirical regularity for how the number of plant and animal species increases with land area [2]—and the scaling of microbial richness with community size [9]. A similar pattern is observed in urban areas—the range of occupations and business types is strongly associated with population [4, 5, 10].

While diversity provides a wide foundation of unique capabilities, specialization refines these capabilities for targeted functions, concentrating on a select few functions or skills to increase performance and competitive edge [11, 12]. For example, certain flowers adapt by developing nectars tailored to their specific pollinators, universities may concentrate on specific research areas like quantum computing, and cities may become hubs for certain sectors such as financial services. Specialization is manifested in the abundance distribution of functions—whether components or resources are concentrated in a few functions or evenly spread across many. Empirically, specialization has been quantified using indices like the Gini coefficient, location quotient, and entropy measures, all of which capture information in the abundance distribution [12–15]. Specialization within the diversification process provides insights into how systems efficiently allocate resources and stand out in competitive environments [16].

Indeed, different systems pursue distinct strategic choices, objectives, and impacts under different specific scopes and subject matters. Accordingly, studies of diversity and specialization have predominantly been siloed within isolated disciplinary boundaries, thus limiting the possibility of a comprehensive understanding. The existing literature in both economics [17, 18] and biology [19–22] often presents diversification and specialization as opposing poles of a linear spectrum. While this perspective is valid, especially in the context of finite resource allocation, it does not universally hold. For instance, research has demonstrated that introducing more minor competitors into a market can, paradoxically, reinforce the dominance of the major players [23]. This indicates that diversification and specialization may be more accurately understood as a two-dimensional problem. In addition, research efforts have often focused on specifics, ranging from the effects of team diversity [1, 24] to the ramifications of organizational diversity [25–27], labor specialization [28, 29], cellular heterogeneity [30–32], or economic diversity and complexity [33, 34]. While these studies are invaluable within their respective fields, findings are often limited to a particular organizational type or the impact of single variables. This segmented approach runs the risk of missing out on identifying overarching principles and theories that underlie the processes of diversification and specialization across a variety of systems.

Thus, there is a need for a comprehensive framework to study diversification and specialization across complex systems. Classically there have been two separate approaches to characterizing commonalities in complex systems: the distributional approach [35, 36] and the feature scaling approach [37–40]. The first tries to connect generative mechanisms to the shape of abundance and probability distributions, the second connects mechanisms to how aggregated properties scale with the overall size of a system often in the form of a power law $Y = Y_0 N^\beta$, where Y is an aggregated quantity, N the system size, and Y_0 a normalization constant. Power-law scaling of aggregated properties is widely argued to be the product of adaptive systems evolving under fixed constraints that yield particular optima [38–40]. Distributional regularities are argued to stem from many processes including self-organized criticality [41, 42], simple growth and attachment rules leading to power-law degree distributions in networks [35], or sampling processes in, for example, ecological systems [43, 44].

Urban systems provide a rich example of how feature scaling relationships have proven helpful for building scientific understanding of complex, highly variable systems. Systematic scaling laws linking socioeconomic outputs and infrastructure costs to population size have inspired models proposing mechanisms including

social network effects [45, 46], geometric and spatial constraints [47], and mobility patterns [48]. Although the scaling exponents of different urban quantities can vary, such differences can arise from common generative principles [46]. Studies have also examined diversity in urban systems using this framework. While new economic functions often originate in large urban centers and subsequently diffuse to smaller ones [49, 50], the diversity of business establishments and occupations exhibits striking regularities across sizes [4, 51].

Beyond urban areas, scaling laws have also been used for cross-system studies. A well-studied regularity is Heaps' law, an empirical pattern originally observed in linguistics in how the number of unique words in a corpus scales with its length, showing a sublinear power law [37]. Similar scaling relationships were later observed in analyses from books in various languages to musical compositions to genomes—using appropriate diversity metrics, they have consistently identified power-law scaling for how number of unique elements scale with size, with exponents between 0.35 and 0.60 [52–57]. The robustness of such empirical scaling laws in novelties within human artifacts has motivated theoretical efforts to uncover their generative mechanisms. Heaps' law is closely related to abundance distributions, as it can be formally derived from Zipf's law [58]. Two main classes of explanations have been proposed—one emphasizing how existing elements open adjacent possibles that enable novel elements to emerge [53, 59], and another focusing on how early commitments constrain subsequent possibilities, limiting what can be added later [60]. Together, these findings suggest that scaling regularities may reflect universal principles governing how complex systems diversify and specialize as they grow. This conceptual foundation motivates our empirical and modeling analysis of how functional diversity and specialization scale with system size across biological and socioeconomic domains.

In this study, we investigate function diversity and specialization in a broad range of socio-economic and biological systems, including bacteria, companies, government agencies, universities, and cities. We explore the interconnected nature of diversification and specialization across a wide range of systems by asking: Do various biological and social systems display common patterns of function diversity and abundance, and to what extent? What general governing principles could determine the function diversity and specialization of complex adaptive systems? We first measure the range (diversity) and concentration of functions (abundance distributions) within each system. We identify functional forms for each attribute to compare disparate systems. We then propose a general model for the growth of function diversity and abundance, explaining both the regularities and differences observed in the empirical data.

Our cross-system perspective recognizes that biological and social systems exhibit common patterns of complexity, adaptation, and evolution. Certain principles, like network dynamics [61] and feedback loops [62], are common across complex systems. Exploring common governing principles across systems may reveal underlying universal patterns that govern complex systems in general. This view is supported by several prior studies that have found commonalities in related domains, such as in the distribution of business types in cities [4, 5], and the abundance fluctuations across biological, ecological, and urban systems [63].

Results

Empirical Results

We define a function as the kind of task an individual component within a system is capable of performing, and function diversity as the number of distinct task types present in the system. This definition is intentionally agglomerated to enable comparisons across structurally and mechanistically diverse systems, ranging from cells to cities. While the internal architectures and generative mechanisms of these systems differ—such as gene regulation in cells, hiring decisions in firms, and curriculum development in universities—we focus on their realized functional repertoires: the observed set of distinct roles or activities that contribute to system-level capabilities.

In biology, proteins are often referred to as the workhorses of the cell because they carry out nearly all of its essential functions. Each type of protein performs a specific task. For example, enzymes catalyze biochemical reactions to produce particular metabolites; motor proteins move substances within the cell; transcription factors are regulatory proteins that control when other proteins are expressed, preventing the simultaneous production of proteins that could interfere with one another. These protein roles have analogs in many
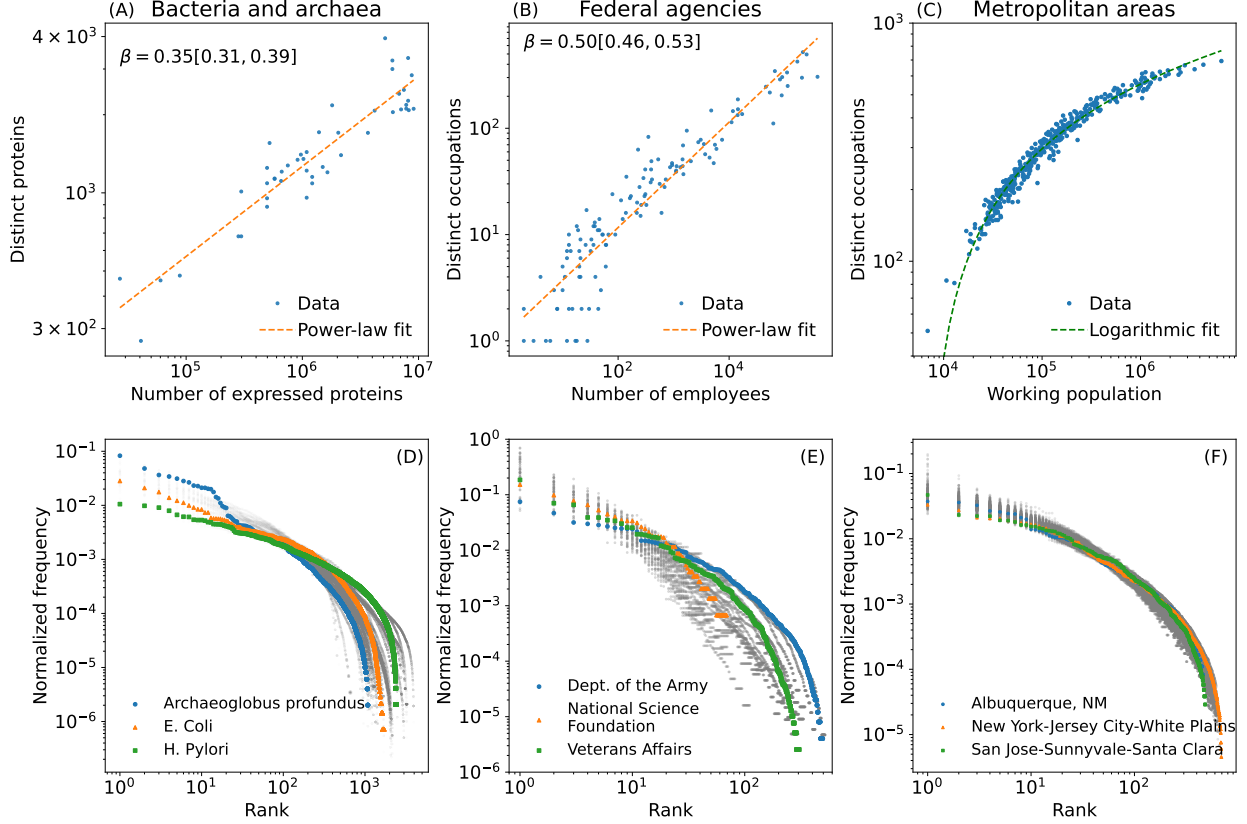
Figure 1: The number of distinct functions versus system size (top row) and rank-frequency distribution of function abundance (bottom row) in several complex adaptive systems. (A,D) Bacterial and archaea cells, (B,E) US federal agencies, and (C,F) US metropolitan statistical areas (MSA). The dependence of the number of functions on size measure N in (A) and (B) can be approximated by the power law, $D \sim N^\beta$. Also shown is the best-fit scaling exponent, β, and its 95% confidence interval in the square brackets. (C) In contrast, logarithmic scaling occurs for MSA's, $D \sim \log N$. In (D–F), grey dots represent all available data points, while select entities are highlighted in color.

socioeconomic systems, such as factory production workers, delivery workers, and managers. Occupations in cities and firms capture the range of economic activities their components can carry out, while academic programs in universities reflect the range of instructional or research activities the institution can perform. Across domains, this functional repertoire reflects the system's capacity to adapt, specialize, and perform complex tasks. By analyzing how function diversity scales with system size, we seek to uncover shared statistical regularities that may point to general principles.

Our study analyzes empirical data from microorganisms to firms to metropolitan areas, including 47 bacteria and archaea cell observations, 125 US federal government agencies, 3,191 observations of Norwegian companies, 2,397 US universities, and 422 US metropolitan areas. For each system, we determine the diversity, the number of distinct functions $D(N)$ and the distribution of these functions' abundance as a function of system size N.

We define system size as the total number of individuals within a system. In social systems, this is simply the number of people in the system, with functions represented by occupations, which are categorized using standardized classifications (see SI for details on data and methods). In bacteria and archaea, key functional components responsible for metabolism and cellular processes are proteins, analogous to workers in an organization. Accordingly, we measure system size in these biological systems as the total number of expressed proteins, while function is quantified as the number of distinct expressed proteins. Figure 1

4

Table 1: Summary of scaling statistics for function diversity in various biological and social complex adaptive systems

	Size measure	Function diversity measure	β	95% CI	R-sq	Obs
Bacteria & archaea	Total expressed protein	Distinct expressed proteins	0.35	[0.31, 0.39]	0.86	47
Norwegian companies	Number of employees	Distinct occupations	0.46	[0.45, 0.46]	0.88	3,191
US federal agencies	Number of employees	Distinct occupations	0.50	[0.46, 0.53]	0.87	125
US universities, associate level	Number of faculty	Distinct academic programs	0.52	[0.50, 0.55]	0.61	1,069
US universities, bachelor level & above	Number of faculty	Distinct academic programs	0.57	[0.55, 0.59]	0.75	1,328
Cities	Working population	Distinct occupations	Logarithmic scaling		0.94	422

illustrates how the range of distinct functions, $D(N)$, scales with the system size N. This scaling framework allows for direct comparisons across diverse systems, as demonstrated in previous studies utilizing scaling analysis [38–40]. The data for bacteria and archaea are visualized in panel A, federal agencies in panel B, and metropolitan areas in panel C. Analysis of additional datasets, including Norwegian companies and universities, are summarized in Table 1, with visualization of scaling relationships included in the SI.

For most systems, this scaling behavior is well-captured by a power-law relation $D = D_0 N^\beta$, where β is the scaling exponent, and D_0 is a constant. For bacteria and archaea, companies, federal agencies, and universities (both associate and bachelor level and above), the scaling exponents are sublinear, ranging between 0.35 to 0.57. This consistency is particularly striking in bacteria and archaea, given that each of the proteomes comes from organisms that have followed unique evolutionary trajectories, have different metabolisms, and occupy unique ecological niches. The sublinear nature of this scaling ($\beta < 1$) further implies that as systems grow in size, the increase in the number of distinct functions is less than proportional. This suggests a tendency for larger systems to reuse or replicate existing functions rather than continuously introducing new ones.

Notably, this behavior closely resembles Heaps' Law, the sublinear power-law scaling for how the number of unique elements scale with size in entities ranging from books in various languages to musical compositions to genomes. While Heaps' Law in human systems is usually associated with the accumulation of novelty in produced artifacts—such as vocabulary in texts or motifs in music—our findings suggest that this same empirical regularity extends to the structure of many human organizations themselves. In other words, the sublinear growth of diversity is not limited to what humans create, but also characterizes many agglomerations they form, such as companies and universities. This shift—from artifacts to collectives—may suggest a deeper principle underlying how complexity scales in systems shaped by human activity.

Among the systems analyzed, a notable exception is metropolitan areas. Unlike other datasets, the scaling of function diversity in metropolitan areas does not follow a power-law pattern but is better captured by a logarithmic relationship, $D = b \log N + c$, as shown in Figure 1(C) (in SI, we show this logarithmic scaling relationship is robust over time). This observation suggests that different mechanisms govern the growth of function diversity in metropolitan areas compared to other systems. Function diversity grows more rapidly with population in smaller cities but decelerates as cities become larger. Notably, the logarithmic scaling observed in metropolitan areas departs from Heaps' Law. This deviation indicates that Heaps-like behavior is not ubiquitous, but instead contingent on structural factors that shape how systems diversify and specialize. If a fundamental underlying mechanism for diversity and specialization exists, it must be capable of explaining both scaling behaviors—under some conditions, Heaps-like behavior emerges, while in

others, diversification and specialization follow alternative patterns.

This logarithmic scaling behavior carries two important implications for urban areas. First, if a city's population grows exponentially with time (t), i.e., $N(t) = N_0 e^{\alpha t}$, where N_0 is the population at time $t = 0$, and α is the growth rate, then function diversity is predicted to grow linearly in time, with $D = \alpha b t + D_0$, where b is the parameter fitted from $D = b \log(N) + c$, and $D_0 = b \log(N_0) + c$ is a constant determined by the initial condition of the system. Then the growth rate of function diversity (dD/dt) is given by the product, αb, which is proportional to the population growth rate. For example, the growth rate α for US urban areas on average is 0.71% in 2019 [64] while the fitted parameter b from data is 128, which together predict that the growth rate for function diversity (dD/dt) is $\alpha b = 0.9$ occupations per year. Second, since most socio-economic metrics in urban areas, such as GDP, wages, and patent production, follow an approximate superlinear power-law relationship with population size [40], our findings suggest that function diversity serves as a key driver of recombinant growth [65, 66]. As functional diversity expands, the combinatorial possibilities for innovation and productivity multiply, contributing to exponential increases in economic output. This underscores the central role of diversity in sustaining and accelerating economic performance in urban systems.

Lastly, we analyze the distribution of function abundance within each organization. Figure 1(D), (E), and (F) illustrate the variation in the relative abundance of functions as a function of rank for bacteria and archaea, federal agencies, and metropolitan areas, respectively. Grey dots represent all data points, while select entities are highlighted in color for clarity.

Despite variation in the most prevalent functions across organizations—such as "nurse" and "medical officer" dominating in Veterans Affairs, while the National Science Foundation's top occupations include "miscellaneous administration and program" and "management and program analysis"—the rank-frequency distributions exhibit a consistent concave shape across all systems. This pattern suggests an exponential decline in function abundance with increasing rank. Notably, the distribution in metropolitan areas displays a striking level of uniformity compared to the more heterogeneous patterns observed in federal agencies and cellular systems. This consistency in urban areas aligns with previous research highlighting similar diversity patterns in business types across cities [4, 10].

Our analysis reveals both commonalities and differences across systems. The commonality is manifested in the fact that within a given system (bacteria, companies, etc.), the range of function diversity generally grows as a sublinear power law with system size. This commonality suggests that while varying in goals, environment, and history, different organisms may share a similar underlying dynamic in how the number of unique functions grows with size. The differences across systems, however, are also important and manifested in two ways. First, the number of unique functions in cities scales differently from the other systems, suggesting the mechanisms that lead to diversity in cities may be fundamentally different from the other entities. Second, among the systems that exhibit power-law scaling, there are variations in the scaling exponents. Our analyses provide the first quantitative assessment enabling comparison across disparate systems.

These findings present both challenges and opportunities for a comprehensive theory that can account for both the observed commonalities and differences. In particular, while existing models have typically predicted either power-law or logarithmic scaling of function diversity (e.g., [4, 59]), none have successfully integrated both patterns. In the next section, we propose a model that is capable of explaining the full spectrum of empirical observations related to function diversity and specialization. By identifying and incorporating key model parameters, our approach aims to shed light on the diverse mechanisms at play, offering a unified understanding of function diversity and specialization across systems.

Mathematical Model

Our model extends the foundational principles of the Yule-Simon model [67, 68], to describe the growth dynamics of a system as new individuals join. Within this framework, a new member that joins the system can lead to one of two scenarios: (i) the new member creates a new function (diversification, with probability denoted as p) or (ii) the newcomer assimilates into a pre-existing function (specialization, with probability $1 - p$). For example, the individuals entering the system are: newly synthesized proteins in a bacteria;

newly hired staff in a federal agency; and a new professional entering the labor market of a city. We assume Markovian dynamics throughout the model, which implies that the probability of adding a new function, or expanding an existing function, is only dependent on the present state of the abundance distributions. Figure 2 visually summarizes this dynamic process.
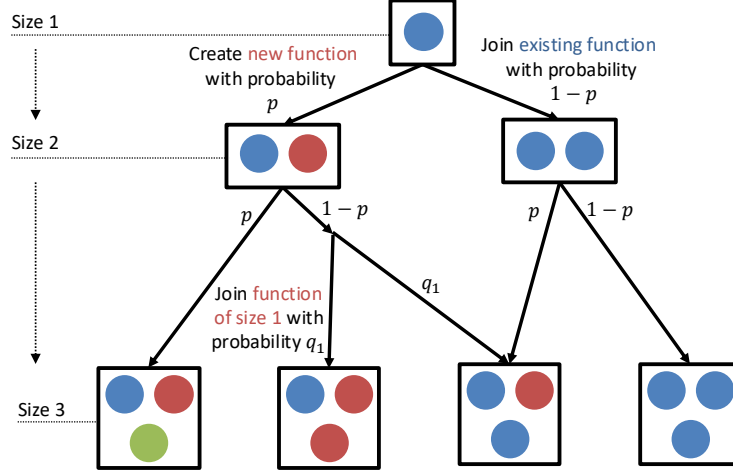

Figure 2: Conceptual diagram of the mathematical model. As each new individual joins the system, it either creates a new function with probability p, where p depends on the existing abundance distribution; or joins an existing function with probability $1 - p$. When it joins an existing function, the probability of joining a particular existing function with k members, q_k, follows a generalized preferential attachment process.

Unlike the classic Yule-Simon model, which assumes a constant rate p of new function creation, our approach introduces a dynamic framework. As a system adapts to a changing environment, new functions may become necessary. However, as the number of existing functions increases, it becomes more likely that the needed functionality is already present. Consequently, the rate of new function creation decreases. When the system contains all possible functions, this rate should approach zero. A simple functional form that would capture the decreasing tendency of p is $p = p_0/D$. The constant p_0 represents effects from environmental factors that are assumed to be the same across systems of the same type.

More generally, the rate of new function creation may not scale inversely with diversity alone, but instead depend on the system's internal structure, as captured by its rank-frequency distribution. A flexible parameterization that accounts for this is,

$$p = \frac{p_0}{\sum_{i=1}^{D} k_i^{\theta}} \,. \tag{1}$$

where k_i denotes the number of individuals within function i. The parameter θ modulates the influence of function abundance on the creation of new functions. When $\theta = 0$, we recover the earlier linear dependency, $p = p_0/D$. For $\theta > 0$, functions with more individuals more strongly suppress the addition of new functions. For $\theta < 0$, the opposite holds. This generalized form allows for a wide range of dynamics to be modeled at the level of individual systems.

When a new individual joins an existing function, we consider the probability of the individual joining each function to be affected by the abundance of the existing functions. Having more individuals in a function can increase the likelihood of a new individual joining this function. In some scenarios, it may have the opposite effect because the existing abundance already adequately meets the system's demands. We can formulate this tradeoff as a nonlinear preferential attachment process—that the probability of an individual joining an existing function with k individuals, q_k, is

$$q_k = \frac{k^{\gamma}}{\sum_{i=1}^{D} k_i^{\gamma}} \,, \tag{2}$$

where γ is a specialization parameter that indicates the extent to which large functions attract newcomers. This equation is similar to previous models of preferential attachment in physics [35, 61, 69] and circular causation [70], positive feedback [71] or agglomeration [72] in economics. Our model introduced γ as a parameter, ranging from 0 to 1, to control the strength of this positive feedback. For example, on the one hand, sublinear preferential attachment [61], that is, $\gamma < 1$, denotes diminishing returns in the effect of function abundance on joining an existing function. On the other hand, the cases $\gamma = 1$ indicate linear preferential attachment, and $\gamma > 1$ corresponds to superlinear preferential attachment, which is extremely sensitive to the existing size of the function.

In summary, our model contains two key parameters, θ and γ, that characterize the effect of *diversification* and *specialization*, respectively. While θ primarily determines the expansion of the function range, γ primarily shapes the accumulation of individuals in functions.

Model predictions and categorization of systems by diversification and specialization parameters

We simulate a system's growth based on Eqs. 1 and 2 and compare the model's predictions with empirical data. The dynamics of the model are shown in Figure 2. Systems of various sizes are treated as points along the model's growth trajectory, representing successive stages in the accumulation of individuals. This assumes that larger systems arise through the sequential addition of individuals to smaller ones. Under this framework, the dynamic model generates cross-sectional predictions that can be directly compared to observed patterns of diversity and abundance across systems of varying sizes.

We then estimate the parameters γ and θ for each system by comparing the function range and abundance generated by simulation with those from data on bacteria, federal agencies, and cities. This estimation is achieved by minimizing the Euclidean distance between normalized rank-frequency distributions in logarithmic-space using the adaptive differential evolution algorithm for optimization (see SI for methods).

Figure 3 (a-c) shows the model's predicted rank-frequency distributions are well aligned with three selected systems: Bartonella henselae (bacterial cell), the Army (federal agency), and Warren-Troy-Farmington Hills (urban area in Michigan). See SI for more examples of rank-frequency distribution predictions. The bottom row of Figure 3 show their respective diversity relationship with those predicted by the model. Both results demonstrate good agreement across all three types of systems for both the rank-frequency distributions and the diversity scaling.

Figure 4 summarizes our calibration results, mapping over a hundred cases onto the parameter space of θ and γ, where each marker represents the calibrated values of one entity, and marker shapes distinguish system type. Systems cluster by type, indicating distinct within-class behavior. The clustering patterns in Figure 4 reveal systematic differences in how functional specialization and diversification unfold across system types. Cities occupy a particularly narrow region of parameter space, centered around $\gamma \approx 0.8$ and $\theta \approx 1$, suggesting a highly consistent balance between diversification suppression and sublinear growth of existing functions. This tight grouping is likely associated with the uniformity in cities' abundance distributions compared to other entities (Fig. 1(F)). In contrast, bacteria and federal agencies span a much broader region of the parameter space, likely reflecting differences in internal constraints and evolutionary or administrative heterogeneity. Together, these results indicate that cities are not only distinct but also unusually consistent in their system-level patterns compared with other entities.

At the same time, patterns emerge across systems along both θ and γ, revealing shared mechanisms that operate both within and across system classes. For instance, bacteria, federal agencies, and cities exhibit parallel trends in the specialization parameter γ, which captures how the abundance of existing functions shapes subsequent growth. All three systems consistently show positive γ values, but predominantly sublinear (below 1), indicating a general tendency toward diminishing returns in specialization. Nevertheless, the variation in γ is wider for federal agencies and bacteria than for cities, suggesting that urban systems may exhibit a more uniform—and possibly universal—specialization pattern.

To test whether the parameter patterns observed for cities are robust to temporal and contextual variation,
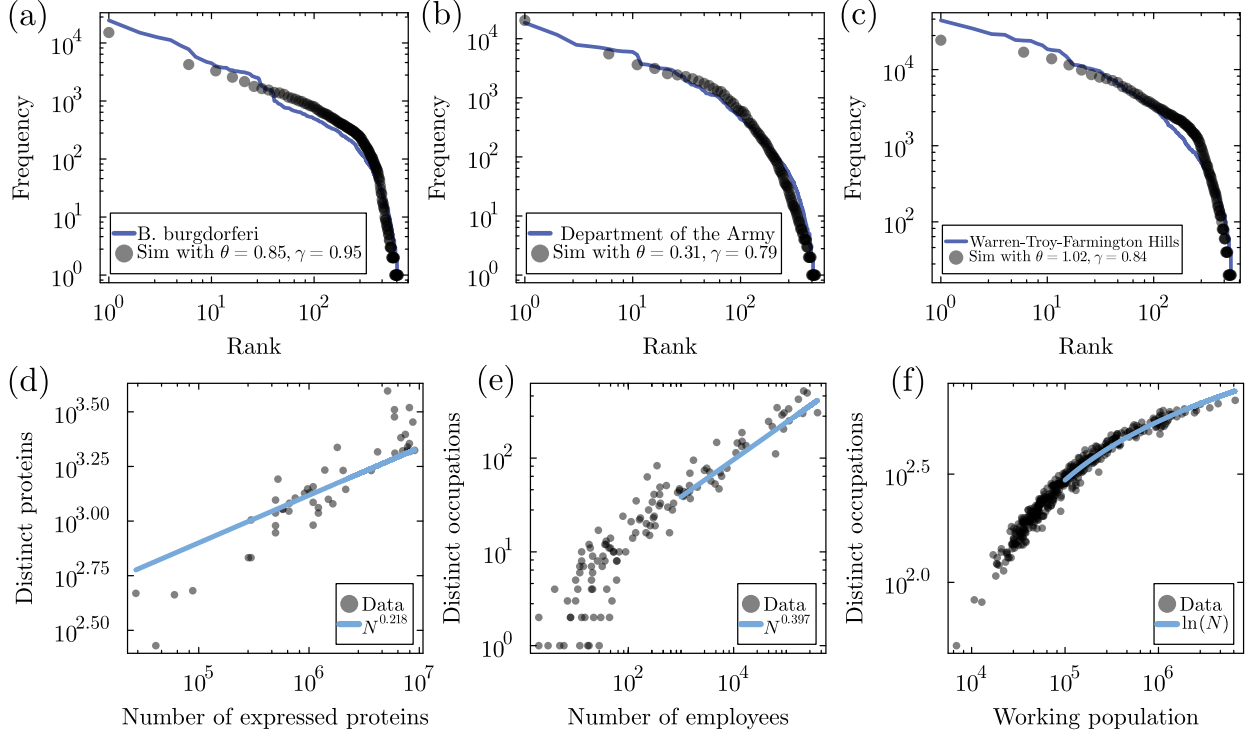
Figure 3: Summary of the model calibration results for the normalized rank-frequency distributions of (a) bacterial cells, (b) federal agencies and (c) cities, each for a given case of that system. The legends show the calibrated values of γ and θ, as well as the particular system in question. Each system was simulated using an initial condition that arose from one of the lesser size organizations in that data set (see SI for further details). Panels (d), (e), and (f) show the respective diversity plots from simulations given the mean values of θ across all calibrations for that system, starting from the initial condition size used for calibration.

we repeated the calibration across multiple years for the four largest urban areas (Figure 5). The results show no systematic temporal trends, despite population of these cities having changed significantly over this time period—indicating that both θ and γ remain stable over time. We also perform robustness tests and demonstrate the stability of these parameters to data noise, initial conditions, and alternative assumptions (see SI).

While the specialization parameter, γ, exhibits common patterns across systems in terms of its mean value, the diversification parameter, θ, differs significantly across systems. It is notably higher for cities than for federal agencies and cells. A higher θ value suggests the existence of a highly abundant function suppresses new function creation. This is in line with the observation that large cities tend to become hubs for certain industries [73], such as San Francisco Bay Area for information technology and the Greater Boston area for biotech. A value of θ around zero implies that an increased range of existing functions, regardless of the size of the functions, contributes towards suppression of new functions. The differences in θ parameters can help explain the differences in scaling behavior observed. When $\theta = 1$, given that $\sum_{i=1}^{D} k_i = N$, we can simplify Eq. 1 to derive $p = p_0/N$. Since $p = dD/dN$, we arrive at the dynamical equation $dD/dN = aD_{\max}/N$. The solution of this equation takes the form of $D \sim \log(N)$, which recovers the logarithmic scaling. Similarly, in the case of $\theta = 0$, we have $\sum_{i=1}^{D} k_i^0 = D$. This leads to $dD/dN \sim 1/D$, which implies the power-law scaling $D \sim N^{1/2}$. For bacteria, Figure 1(A) shows that diversification happens faster than in cities, but slower than in federal agencies, and correspondingly the mean calibrated value of θ is between 0 and 1.

Each class of systems occupies distinct regions of parameter space. First, the values of θ found for cities imply that diversity growth is most impeded when large functions are abundant. Second, for federal agencies, smaller θ values imply that functions more equitably impede diversity growth. Finally, for bacterial cells,
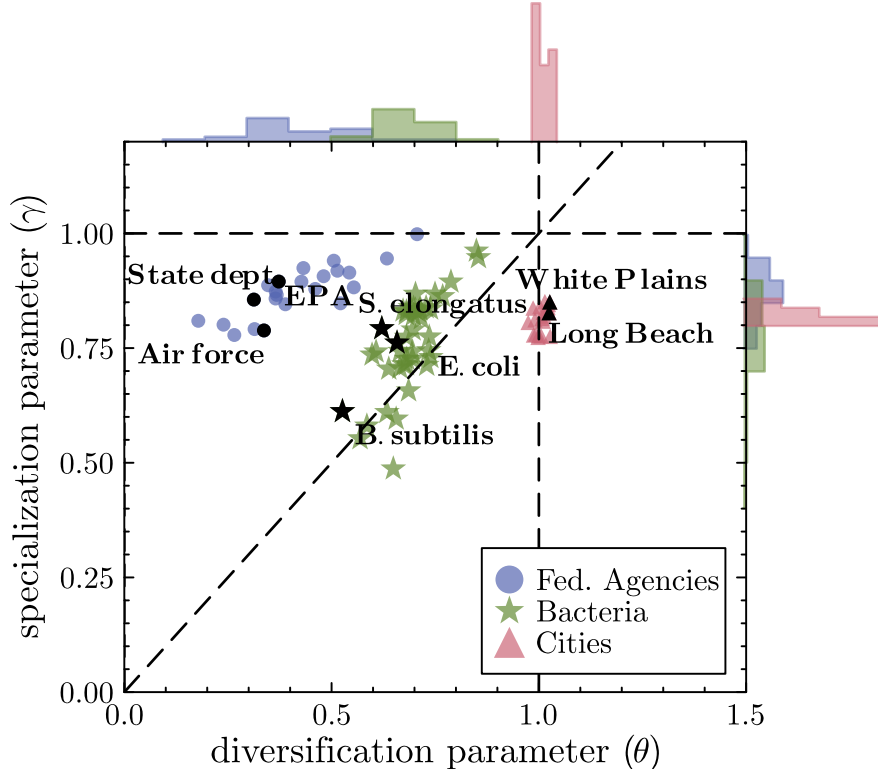
9

Figure 4: The parameter space of specialization (γ) and diversification (θ) estimated for each system across different classes of systems. The vertical dashed line shows $\theta = 1$, the horizontal dashed line shows $\gamma = 1$, and the diagonal dashed line shows $\gamma = \theta$. We selected the federal agencies with $N > 1000$ (40 instances), all of the bacterial data we had access to (46 species), and the largest twenty cities.

an intermediate θ value implies an in-between level of diversity growth still impeded by large functions. Cities express the largest value of θ (approximately 1) that simultaneously allows for the maximization of concentration in a few functions while maintaining consistent rank-frequency distributions without leading to gelation behavior that occurs for $\gamma > 1$ [61]. In this sense, gelation provides a natural bound to function growth since it does not allow for the robustness or diversity necessary for organizations to survive complex environments.

By enabling cross-domain comparisons, we identify important commonalities and differences across complex systems. The barrier to introducing new functions varies across system types. Yet, once a function is added, it tends to grow following a common sublinear preferential attachment dynamics. The variations may stem from differences in system goals and structures. Federal agencies and bacterial cells function as cohesive organisms with well-defined boundaries, where adaptation occurs at the level of the whole entity. In federal agencies, the introduction of new occupations is typically decided in a centralized manner. Cities, on the other hand, lack unified objectives and resemble ecosystems: decentralized, and adaptation occurs at the level of the individuals. The introduction of new occupational categories is decided in a decentralized fashion. For organism-like systems, it is beneficial to develop a full set of functional capabilities to effectively adapt to environmental challenges. Thus, even when some functions dominate, there remains a drive to cover all essential bases. In cities, however, large functions may inhibit the emergence of others due to competitive or self-reinforcing dynamics. Once functions are introduced into these systems, the growth of function abundance follows a shared pattern across all systems: sublinear preferential attachment. This suggests that while the drivers of function *creation* differ based on system architecture and goals, the *growth* of functions once introduced follows a remarkably universal process.
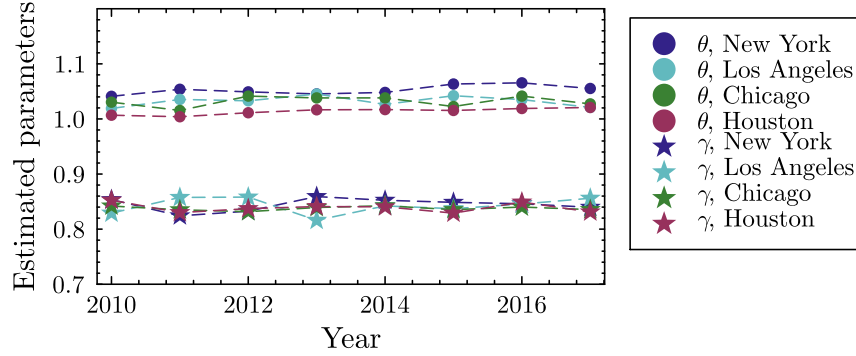
Figure 5: Calibrated parameters (θ and γ) for the four largest metropolitan areas in the longitudinal dataset, by year. Parameter values show no systematic temporal trends.

Discussion

We identify common patterns in the diversity and abundance of functions across a broad spectrum of biological and social complex systems. In systems such as bacteria, federal agencies, companies, and universities, function diversity can be approximated by a power-law relationship with size, characterized by sublinear exponents ranging from 0.35 to 0.57, depending on the system. This is consistent with Heaps' Law. However, cities differ significantly from these systems and exhibit logarithmic scaling in function diversity and greater commonality in abundance distributions. This empirical anomaly may stem from cities' fundamentally different structures and goals.

To reconcile these empirical patterns, we propose a general model for the dynamics of function diversity and abundance. The model is characterized by two fundamental parameters: a diversification parameter, which governs the expansion of the function range, and a specialization parameter, which grows the function concentration. We estimate both parameters from the data for bacterial cells, federal agencies, and cities. Notably, we observe consistent sublinear preferential attachment across all three systems, with generally small variations in the specialization parameter. However, the diversification parameter varies significantly across systems, suggesting different drivers for function range expansion. We emphasize that this universality lies at the level of generative mechanism rather than identical empirical outcomes. For simplicity, the model that we propose is fully Markovian, although future extensions to non-Markovian models with memory [74, 75] would be a beneficial addition to the literature, especially in the context of memory of organizational routines. The tight clustering of cities in parameter space implies a high degree of structural regularity in how urban systems specialize and diversify. This consistency supports the view that urban economies may be shaped by shared underlying mechanisms, as proposed in the urban scaling literature [76]. Extending this analysis to cities across different geographical and governance contexts would provide a useful test of the model's generality.

Our framework offers a complementary alternative to several influential theories in biology and ecology that have sought to explain the origins and patterns of diversity. Classical models such as the Island Biogeography Theory [77], the Unified Neutral Theory of Biodiversity and Biogeography (UNTBB) [43, 78], and Maximum Entropy [44] approaches have all been pivotal in modeling biodiversity through the lens of ecological equilibria, spatial constraints, or information-theoretic principles. In particular, the UNTBB has offered a controversial but widely used null model, helping researchers benchmark observed diversity patterns across ecosystems and even genomes. Unlike these models, which are typically grounded in ecological or evolutionary narratives specific to biological systems, our approach derives from a generalized process-based model that incorporates both diversification and specialization dynamics, and it applies equally well to socioeconomic systems such as cities and federal agencies. This broader scope enables cross-domain comparisons and identifies shared underlying mechanisms, while also capturing key differences—such as the logarithmic versus power-law scaling of diversity—in a unified parameter space.

A key to this cross-domain comparison is the definition of diversity. Our use of the term in this paper specifically refers to the number of unique functions in the sub-components within a system. This definition differs from entropy-based measures stemming from information theory, which have also been used as a diversity index in the biological and social systems literature. Entropy-based measures combine two properties of distributions—richness and evenness—into a single index. While valuable in many contexts, entropy-based measures have the limitation of possessing mathematical properties that do not align with theoretical or intuitive understandings of diversity in applied domains [79]. In contrast, our diversity metric solely measures richness, the system's range of unique capabilities. We explicitly separate the dimension of evenness, which we refer to as specialization, into a distinct second axis.

Our analysis relies on classification systems that delineate what constitutes a distinct function in each domain, which have limitations. For example, a protein's role can sometimes vary depending on the cellular context across species, most retain consistent biochemical functions that permit meaningful functional classification. Social systems exhibit similar variations within function categories: occupations are defined by shared core tasks even as their implementation varies across organizations. The "chief executive" occupation category, for example, contains leaders of both start-ups and multinational corporations, whose daily activities differ, yet they are grouped together because they share core tasks of strategic oversight and coordination. The protein–occupation analogy thus reflects commonality in functional organization across diverse contexts rather than a literal one-to-one correspondence. Admittedly, classification systems, especially those in social systems, are shaped by administrative needs. However, they are updated regularly to reflect evolving labor and knowledge structures, and are widely used in real-world economic and policy decisions. We use the most granular level of these taxonomies to approximate a system's repertoire of task specializations. Though no classification is perfect, the persistence of scaling patterns across domains that use different classification schemes suggests that the underlying dynamics of functional diversification and specialization reflect deeper regularities in system organization despite limitations in the classification systems.

Although proteins can exhibit context-dependent roles across species, most retain consistent biochemical functions that permit meaningful functional classification. Social systems show a comparable pattern: occupations are defined by shared core tasks even as their implementation varies across organizations. A start-up founder and a corporate CEO, for example, differ in daily activities yet share core functions of coordination and strategic control, and are thus categorized under the same SOC category. The protein–occupation analogy thus reflects a common structure of functional organization across diverse contexts rather than a literal one-to-one correspondence

Our work provides a parsimonious framework for unifying observed patterns across domains, but naturally abstracts away many details. This opens several directions for extension. While our model accounts for cross-sectional differences in scaling behavior across system types, future work could extend this framework to incorporate temporal dynamics, considering the evolution of the diversification and specialization parameters (θ, γ) over time at different stages in system development. Our proposed framework represents an effective process, but does not explicitly capture the causal dynamics through which new functions or components emerge. Future work could extend this framework by explicitly modeling the structural dependencies among functions through hierarchical trees or network-based mechanisms. These include hierarchical specialization (depth), recombination across existing types (horizontal linkage), and the introduction of entirely novel types (breadth). Another promising extension involves accounting for function obsolescence. Our model focuses on the addition of individuals and functions, but individuals also leave organizations and cities, just as genes can naturally be removed from genomes through evolution. Entire functions can also become obsolete, such as the obsolescence of switchboard operators who manually connected telephone calls. Although the mechanisms of agent removal and function obsolescence are still areas of active research [80, 81], it could be useful for future research to consider the removal of individuals and functions in complex systems and how it would affect diversification and specialization. Lastly, our model treats the introduction of new functions as a single process, without distinguishing between the creation of globally novel functions and the adoption of ones that exist in other entities. In both biological and social systems, new functions arise within an entity and spread through diffusion—for example, via horizontal gene transfer in biological systems [82] and imitation in social systems [83]—reflecting a shared innovation–diffusion dynamic. Because the adoption of functions that exist in other organisms is likely far more common than the creation of globally new ones,

it may dominate the observed empirical patterns. Future work could explicitly separate these processes to assess their relative contributions.

Our work makes several contributions to the literature. First, in contrast to the prevalent trend of studying diversification and specialization within specific system types, our work presents an integrated analysis across a diverse array of systems, encompassing both biological (bacteria) and social systems (companies, government agencies, universities, cities). This approach identifies commonalities across many biological and social systems, while also identifying notable exceptions. It provides testable hypotheses to explore patterns in other systems. Second, while previous models explained power-law or logarithmic scaling relationships using distinct formulations and assumptions, we develop a general model that explains both types of scaling within one unifying framework. Third, by estimating parameters in the framework of this mathematical model, we can quantify common characteristics across complex systems and evaluate how systems differ from each other.

Code availability

The data and code necessary to replicate the results of this study are available at https://github.com/jamesholehouse/Calibrations-for-Function-Diversity.

Author contributions

V.C.Y. and J.H. contributed equally to this work. Data gathering and empirical analysis: V.C.Y., J.I.A.; Model development, simulation, and calibration: J.H., S.R.; Writing – Initial Draft: V.C.Y, J.H.; Writing – Review & Editing: All authors; Conceptualization: All authors.

Acknowledgments

References

1. Page, S. E. *The Difference: How the power of diversity creates better groups, firms, schools, and societies* (Princeton University Press, 2008).

2. Whittaker, R. H. EVOLUTION AND MEASUREMENT OF SPECIES DIVERSITY. *Taxon* **21,** 213–251 (1972).

3. Quigley, J. M. Urban Diversity and Economic Growth. *Journal of Economic Perspectives* **12,** 127–138 (June 1998).

4. Youn, H. *et al.* Scaling and universality in urban economic diversification. *Journal of the Royal Society Interface* **13,** 20150937 (2016).

5. Luis M. A, B., Horacio, S. & Hyejin, Y. Professional diversity and the productivity of cities. *Scientific Reports* **4** (2014).

6. Kline, M. A. & Boyd, R. Population size predicts technological complexity in Oceania. *Proceedings of the Royal Society B: Biological Sciences* **277,** 2559–2564 (2010).

7. Simpson, E. Measurement of Diversity. *Nature* **163** (1949).

8. Kogut, B. & Zander, U. Knowledge of the firm, Combinative Capabilities, and the Replication of Technology. *Organization Science* **3,** 383–397. ISSN: 10477039, 15265455. http://www.jstor.org.proxy.cc.uic.edu/stable/2635279 (Oct. 1992).

9. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences* **113,** 5970–5975 (2016).

10. Hong, I., Frank, M. R., Rahwan, I., Jung, W. S. & Youn, H. The universal pathway to innovative urban economies. *Science Advances* **6,** 1–7. ISSN: 23752548 (2020).

11. Perloff, H., E. Dunn, E. L. & Muth, R. *Regions, Resources, and Economic Growth* (Baltimore, MD: John Hopkins Press, 1960).

12. Kim, S. Expansion of Markets and the Geographic Distribution of Economic Activities: The Trends in U. S. Regional Manufacturing Structure, 1860-1987. *The Quarterly Journal of Economics* **110,** 881–908 (1995).

13. Frank, M. R., Sun, L., Cebrian, M., Youn, H. & Rahwan, I. Small cities face greater impact from automation. *Journal of the Royal Society Interface* **15,** 139 (2018).

14. Henderson, J. V. *Urban Development: Theory, Fact, and Illusion OUP Catalogue* **9780195069020.** https://ideas.repec.org/b/oxp/obooks/9780195069020.html (Oxford University Press, 1991).

15. Palan, N. Measurement of Specialization: The Choice of Indices. *FIW Working Paper* **62** (2010).

16. Page, S. in *Diversity and complexity* (Princeton University Press, 2010).

17. Beaudry, C. & Schiffauerova, A. Who's right, Marshall or Jacobs? The localization versus urbanization debate. *Research policy* **38,** 318–337 (2009).

18. Imbs, J. & Wacziarg, R. Stages of diversification. *American Economic Review* **93,** 63–86 (2003).

19. MacArthur, R. & Levins, R. The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist* **101,** 377–385 (1967).

20. Cracraft, J. Biological diversification and its causes. *Annals of the Missouri Botanical Garden* **72,** 794–822 (1985).

21. Mahler, D. L., Ingram, T., Revell, L. J. & Losos, J. B. Exceptional convergence on the macroevolutionary landscape in island lizard radiations. *Science* **341,** 292–295 (2013).

22. Day, E., Hua, X. & Bromham, L. Is specialization an evolutionary dead end? Testing for differences in speciation, extinction and trait transition rates across diverse phylogenies of specialists and generalists. *Journal of Evolutionary Biology* **29,** 1257–1267 (2016).

23. Chu, J. S. A theory of durable dominance. *Strategy Science* **3,** 498–512 (2018).

24. Modi, N., Lungeanu, A. & DeChurch, L. e. a. The differential impacts of team diversity as variability versus atypicality on team effectiveness. *Sci Rep* **15** (2025).

25. Palich, L. E., Cardinal, L. B. & Miller, C. C. Curvilinearity in the diversification–performance linkage: an examination of over three decades of research. *Strategic Management Journal* **21,** 155–174 (2000).

26. Horwitz, S. K. in, 329 (2015).

27. Milliken, F. J. & Martins, L. L. Searching for Common Threads: Understanding the Multiple Effects of Diversity in Organizational Groups. *Academy of Management Review* **21,** 402–433 (1996).

28. Becker, G. S. & Murphy, K. M. The Division of Labor, Coordination Costs, and Knowledge. *The Quarterly Journal of Economics* **107,** 1137–1160. ISSN: 00335533, 15314650. http://www.jstor.org/stable/2118383 (2022) (1992).

29. Hosseinioun, M., Neffke, F., Zhang, L. & Youn, H. Skill dependencies uncover nested human capital. *Nature Human Behavior* **9,** 673–687 (2025).

30. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297,** 1183–1186 (2002).

31. Holehouse, J., Cao, Z. & Grima, R. Stochastic modeling of autoregulatory genetic feedback loops: A review and comparative study. *Biophysical Journal* **118,** 1517–1525 (2020).

32. Altschuler, S. J. & Wu, L. F. Cellular heterogeneity: do differences make a difference? *Cell* **141,** 559–563 (2010).

33. Hausmann, R., Hidalgo, C. A., Bustos, S., Coscia, M. & Simoes, A. *The Atlas of Economic Complexity: Mapping Paths to Prosperity* (MIT Press, 2014).

34. Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences* **106,** 10570–10575 (2009).

35. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286,** 509–512 (1999).

36. Broido, A. D. & Clauset, A. Scale-free networks are rare. *Nature Communications* **10,** 1017 (2019).

37. Heaps, H. S. *Information Retrieval, Computational and Theoretical Aspects* (Academic Press, 1978).

38. West, G. B. *Scale: The Universal Laws of Life, Growth, and Death in Organisms, Cities, and Companies* (Penguin, 2018).

39. West, G. B. The origin of universal scaling laws in biology. *Physica A: Statistical Mechanics and its Applications* **263,** 104–113 (1999).

40. Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences* **104,** 7301–7306 (2007).

41. Bak, P., Tang, C. & Wiesenfeld, K. Self-organized criticality: An explanation of the 1/f noise. *Physical review letters* **59,** 381 (1987).

42. Bak, P., Tang, C. & Wiesenfeld, K. Self-organized criticality. *Physical review A* **38,** 364 (1988).

43. Hubbell, S. P. in *The unified neutral theory of biodiversity and biogeography (MPB-32)* (Princeton University Press, 2011).

44. Harte, J. *Maximum entropy and ecology: a theory of abundance, distribution, and energetics* (OUP Oxford, 2011).

45. Bettencourt, L. M. A. The origins of scaling in cities. *Science* **340,** 1438–1441 (2013).

46. Yang, V. C., Papachristos, A. V. & Abrams, D. M. Modeling the origin of urban-output scaling laws. *Physical Review E* **100,** 032306 (2019).

47. Molinero, C. & Thurner, S. How the geometry of cities determines urban scaling laws. *Journal of the Royal Society interface* **18,** 20200705 (2021).

48. Ribeiro, F. L., Meirelles, J., Ferreira, F. F. & Neto, C. R. A model of urban scaling laws based on distance dependent interactions. *Royal Society Open Science* **4,** 160926 (2017).

49. Pumain, D., Paulus, F., Vacchiani-Marcuzzo, C. & Lobo, J. An evolutionary theory for interpreting urban scaling laws. *Cybergeo: European Journal of Geography* (2006).

50. Hong, I., Frank, M. R., Rahwan, I., Jung, W.-S. & Youn, H. The universal pathway to innovative urban economies. *Science Advances* **6,** eaba4934 (2020).

51. Bettencourt, L. M. A., Samaniego, H. & Youn, H. Professional diversity and the productivity of cities. *Scientific reports* **4,** 5393 (2014).

52. Mazzolini, A., Colliva, A., Caselle, M. & Osella, M. Heaps' law, statistics of shared components, and temporal patterns from a sample-space-reducing process. *Physical Review E* **98,** 052139 (2018).

53. Tria, F., Loreto, V., Servedio, V. D. P. & Strogatz, S. H. The dynamics of correlated novelties. *Scientific Reports* **4,** 1–8 (2014).

54. Benz, R. W., Swamidass, S. J. & Baldi, P. Discovery of power-laws in chemical space. *Journal of Chemical Information and Modeling* **48,** 1138–1151 (2008).

55. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: The bacterial pan-genome. *Current Opinion in Microbiology* **11,** 472–477 (2008).

56. Park, S.-C., Lee, K., Kim, Y. O., Won, S. & Chun, J. Large-scale genomics reveals the genetic characteristics of seven species and importance of phylogenetic distance for estimating pan-genome size. *Frontiers in Microbiology* **10,** 834 (2019).

57. Serra-Peralta, M., Serrà, J. & Corral, Á. Heaps' law and vocabulary richness in the history of classical music harmony. *EPJ Data Science* **10,** 40 (2021).

58. Van Leijenhorst, D. C. & van der Weide, T. P. A formal derivation of Heaps' Law. *Information Sciences* **170,** 263–272 (2005).

59. Loreto, V., Servedio, V. D., Strogatz, S. H. & Tria, F. in *Creativity and Universality in Language* 59–83 (Springer, 2016).

60. Corominas-Murtra, B., Hanel, R. & Thurner, S. Understanding scaling through history-dependent processes with collapsing sample space. *Proceedings of the National Academy of Sciences* **112,** 5348–5353 (2015).

61. Krapivsky, P. L. & Redner, S. Organization of growing random networks. *Physical Review E* **63,** 066123 (2001).

62. Alon, U. Network motifs: Theory and experimental approaches. *Nature Reviews Genetics* **8,** 450–461 (2007).

63. George, A. B. & O'Dwyer, J. Universal abundance fluctuations across microbial communities, tropical forests, and urban populations. *Proceedings of the National Academy of Sciences* **120,** e2215832120 (2023).

64. The World Bank. *Urban Population Growth – United States.* `https://data.worldbank.org/indicator/SP.URB.GROW?locations=US` 2019.

65. Weitzman, M. L. Recombinant Growth. *The Quarterly Journal of Economics* **113,** 331–360. ISSN: 0033-5533. eprint: `https://academic.oup.com/qje/article-pdf/113/2/331/5403892/113-2-331.pdf`. `https://doi.org/10.1162/003355398555595` (May 1998).

66. Youn, H., Strumsky, D., Bettencourt, L. M. A. & Lobo, J. Invention as a combinatorial process: evidence from US patents. *Journal of The Royal Society Interface* **12,** 20150272. eprint: `https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2015.0272`. `https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2015.0272` (2015).

67. Yule, G. U. A mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. Series B* **213,** 21–87 (1925).

68. Simon, H. A. On a class of skew distribution functions. *Biometrika* **42,** 425–440 (1955).

69. Krapivsky, P. L., Rodgers, G. J. & Redner, S. Degree distributions of growing networks. *Physical Review Letters* **86,** 5401 (2001).

70. Myrdal, G. & Sitohang, P. *Economic theory and under-developed regions* (Duckworth London, 1957).

71. Arthur, W. B. Competing Technologies, Increasing Returns, and Lock in by Historical Events. *Economic Journal* **99,** 92–99 (1990).

72. Hoover, E. M. *The Location of Economic Activity* (New York: McGraw-Hill, 1948).

73. Balland, P.-A. *et al.* Complex economic activities concentrate in large cities. *Nature Human Behaviour* **4,** 248–254 (2020).

74. Ghorbani, M., Jonckheere, E. A. & Bogdan, P. Gene expression is not random: scaling, long-range cross-dependence, and fractal characteristics of gene regulatory networks. *Frontiers in physiology* **9,** 1446 (2018).

75. Reed, E. A., Ramos, G., Bogdan, P. & Pequito, S. The role of long-term power-law memory in controlling large-scale dynamical networks. *Scientific Reports* **13,** 19502 (2023).

76. Bettencourt, L. M. A. *Introduction to urban science: evidence and theory of cities as complex systems* (MIT press, 2021).

77. MacArthur, R. H. & Wilson, E. O. *The theory of island biogeography* (Princeton university press, 2001).

78. Serra, F., Becher, V. & Dopazo, H. Neutral theory predicts the relative abundance and diversity of genetic elements in a broad array of eukaryotic genomes. *PLoS One* **8,** e63915 (2013).

79. Jost, L. Entropy and diversity. *Oikos* **113,** 363–375 (2006).

80. Lee, E. D. *et al.* Synthesis of innovation and obsolescence. *arXiv preprint arXiv:2505.05182* (2025).

81. Saavedra, S., Reed-Tsochas, F. & Uzzi, B. Asymmetric disassembly and robustness in declining networks. *Proceedings of the National Academy of Sciences* **105,** 16466–16471 (2008).

82. Arnold, B. J., Huang, I.-T. & Hanage, W. P. Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology* **20,** 206–218 (2022).

83. DiMaggio, P. J., Powell, W. W., *et al.* The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American sociological review* **48,** 147–160 (1983).