scientific reports

Check for updates

OPEN

Scarcity of scale-free topology is universal across biochemical networks

Harrison B. Smith[1,5], Hyunju Kim[1,2,3] & Sara I. Walker[1,2,3,4]✉

Biochemical reactions underlie the functioning of all life. Like many examples of biology or technology, the complex set of interactions among molecules within cells and ecosystems poses a challenge for quantification within simple mathematical objects. A large body of research has indicated many real-world biological and technological systems, including biochemistry, can be described by power-law relationships between the numbers of nodes and edges, often described as "scale-free". Recently, new statistical analyses have revealed true scale-free networks are rare. We provide a first application of these methods to data sampled from across two distinct levels of biological organization: individuals and ecosystems. We analyze a large ensemble of biochemical networks including networks generated from data of 785 metagenomes and 1082 genomes (sampled from the three domains of life). The results confirm no more than a few biochemical networks are any more than super-weakly scale-free. Additionally, we test the distinguishability of individual and ecosystem-level biochemical networks and show there is no sharp transition in the structure of biochemical networks across these levels of organization moving from individuals to ecosystems. This result holds across different network projections. Our results indicate that while biochemical networks are not scale-free, they nonetheless exhibit common structure across different levels of organization, independent of the projection chosen, suggestive of shared organizing principles across all biochemical networks.

Statistical mechanics was developed in the nineteenth century for studying and predicting the behavior of systems with many components. It has been hugely successful in its application to those physical systems well-approximated by idealized models of non-interacting particles. However, real-world systems are often much more complex, leading to a realization over the last several decades that new statistical approaches are necessary to describe biological and technological systems. Among the most natural mathematical frameworks for developing the necessary formalism is network theory, which projects the complex set of interactions composing real systems onto an abstract graph representation[1–7]. Such representations are powerful in their capacity to quantitatively describe the relationship between components of complex systems and because they permit inferring function and dynamics from structure[8–12].

Network theory has been especially useful for studying metabolism. Metabolism consists of catalyzed reactions that transform matter along specific pathways, creating a complex web of interactions among the set of molecular species that collectively compose living things[13–17]. It is the collective behavior of this system of reactions that must be understood in order to fully characterize living chemical processes—counting only individual components (molecules) is inadequate. The structure of how those components interact with one another (via reactions) really matters: in fact it is precisely what separates organized biological systems from messy chemical ones[18–20].

Within the formalism of network theory, one of the simplest ways to capture insights into the global structure of a network is to analyze the shape of its degree distribution. A huge volume of research into various complex biological, technological and social networks has therefore focused on identifying the scaling behavior of the corresponding degree distributions for network projections describing those systems. One of the most significant results emerging from these analyses is that many networks describing real-world systems exhibit ostensibly "scale-free" topology[21–25], characterized by a power-law degree distribution. The allure of scale-free networks

[1]School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA. [2]Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe, AZ, USA. [3]ASU-SFI Center for Biosocial Complex Systems, Arizona State University, Tempe, AZ, USA. [4]Santa Fe Institute, Santa Fe, NM, USA. [5]Present address: Earth-Life Science Institute, Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan. ✉email: sara.i.walker@asu.edu

is in part driven by the simplicity of their underlying generative mechanisms, for example a power-law degree distribution can be produced by relatively simple preferential attachment algorithms[21], or to a lesser extent through optimization principles[26]. For truly scale-free networks the probability to find a node with degree x should scale as:

$$f(x) = x^{-\alpha} . \tag{1}$$

For numerous biological and technological systems, including metabolic networks, the scaling exponent, α, is reported with values in the range $2 < \alpha < 3$. The apparent ubiquity of scale-free networks across biological, technological and social networks has fueled some to conjecture scale-free topology as a unifying framework for understanding all such systems, with the enticing possibility these seemingly diverse examples could in reality arise from relatively simple, universal generating mechanisms[21,25–28].

However, this story is far from complete. Recently, Broido and Clauset developed statistical tests to rigorously examine whether observed distributions share characteristics with a power-law, or are instead more similar to other heavy tailed distributions, and have revealed that true scale-free networks may not be as ubiquitous as previously supposed[29,30]. These tests reveal that while it is superficially possible for a network to appear scale-free, more rigorous analysis can reveal a structure more similar to other heavy-tailed distributions such as the log-normal distribution, or even non heavy-tailed distributions like the exponential distribution[28–31].

The problem of characterizing the global structure of real-world systems is further compounded by the fact there are often many ways to coarse-grain a real system to generate a network representation, each corresponding to a different way for set of interactions to be projected onto a graph. For example, metabolic networks may be represented as unipartite or bipartite graphs, depending on whether one chooses to focus solely on the statistics over molecules (or reactions) and their interactions (requiring a unipartite representation) or instead to include both molecules and reactions as explicit nodes in the graph (where molecules and reactions represent two classes of nodes in a bipartite representation)[32–34]. These graphs can have different large-scale topological properties, even when projected from the same underlying system. This raises the question of determining which projection to analyze, and whether or not a real-world system should be considered "scale-free" if only some of its network projections exhibit power-law degree distributions. In Broido and Clauset's classification, "scale freeness' is assessed ranging from "Not scale-free" to "strongest"[30]. Their approach provides methods for statistically analyzing the different network projections of real-world systems to determine how well scale-free structure can describe the properties of the high-dimensional underlying system, when it is projected into lower dimensional, coarse-grained network representations.

The goal of assessing different network projections in order to classify "scale freeness" is intended to be as thorough as possible in identifying relevant features of a systems' complex network structure. However, there is debate about whether this criterion is too strict. In particular, some researchers have argued that depending on the system being analyzed, it may not make sense to represent and equally weigh many different network projections (see e.g. debate by Barabasi and others[35]). Herein, we aim to be as agnostic as possible about which projections are best suited for capturing how the many components of biochemical systems interact, as this is an open question in its own right. Given our aim to broadly assess the scaling of biochemical systems, we therefore follow Broido and Clauset and consider all possible projections available from the underlying data. A second potential criticism of this approach concerns whether or not it matters if long-tail networks are not precisely scale-free. Our goal here is to report the statistical properties of biochemical networks across scales and agree in some contexts it might not matter if they are precisely scale-free, but in others it might. For example, many different heavy tailed distributions share the property of having a high degree-squared mean, $\langle k^2 \rangle$, and in many applications this indicates a high-robustness to failure[24,36–38]. Although these networks may share some properties, they can also have very different underlying generative mechanisms. There has even been recent work proposing that the phrase "scale-free" should be used only for networks generated using the preferential-attachment mechanism[39]. Since we do not yet know the generative mechanisms that best explain the structure of biochemical networks, our goal herein is to provide the most rigorous classification of their structure that might enable future research in this area.

The novelty in our approach is recognizing that in order to understand the structure of real-world biological (and technological) systems, the relevant organizational level for performing analyses must also be considered. In particular, biological and technological systems are often hierarchical in their organization, with interactions across multiple levels. In example, while it is possible to study the biochemistry of individual species, ultimately the function of species in natural systems depends on the complex interplay of interactions among the many species within an ecosystem. Indeed, universal properties of life are now recognized to be characterized at the scale of ecosystems as much as they are at the scale of individual organisms[40,41].

In what follows, we analyze a large set of biochemical systems including data from 785 metagenomes (ecosystem-level) and 1082 genomes (individual-level, sampled from each of the three domains of life). Our results include the first analysis of scale-free network structure for the different projections of ecosystem-level biochemistry, significantly expanding on earlier work focusing on the large-scale structure of individual metabolic networks only[13,29,30,32–34]. Like Broido and Clauset, we consider all possible projections of biochemical systems to graphs simultaneously, whereas most prior work on the organization of biochemistry has only considered one or at most a few projections[17,42–45]. We find a majority of biochemical networks are not scale-free, independent of projection or level of organization. We also demonstrate how the network properties analyzed herein can be used to distinguish individual and ecosystem level networks, and find that independent of projection, individuals and ecosystems share very similar structure. These results have potentially deep implications for identifying underlying rules of biochemical organization at both the individual and ecosystem-level by providing constraints

on whether the same or different generative mechanisms could operate to organize biochemistry across multiple scales.

Results

We use the statistical methods developed by Broido and Clauset[30] in what follows. All identified biochemical reactions encoded in each genome and metagenome were used to construct eight distinct network representations for each. This resulted in 8656 total network projections across the 1082 genome-level biochemical datasets, and 6280 total network projections across the 785 metagenome-level biochemical datasets. Each representation can be viewed as a different coarse-graining of the underlying system of reactions (i.e. the underlying dataset) (Fig. 1). We determine whether or not these datasets are scale-free, and analyze the aspects of them, and their diverse projections, that tend to lend themselves to be more or less scale-free. The alternative distributions that we compare to the power-law are: The exponential distribution, the log-normal distribution, the stretched exponential distribution, and the power-law distribution with a cutoff (see[29,30] for more details on these distributions).

We first classified each dataset in terms of how scale-free it is. Data are classified as: *Super-Weak* when for $\geq$ 50% of network projections, no alternative distributions are favored over power-law; *Weakest* if for $\geq$ 50% of network projections, a power-law cannot be rejected ($p \geq 0.1$); *Weak* if it meets the requirements for Weakest, and there are $\geq$ 50 nodes in the distribution's tail ($n_{\text{tail}} \geq 50$); *Strong* if it meets the requirements of both Super-Weak and Weak, and the median scaling exponent satisfies ($2 < \hat{\alpha} < 3$); and *Strongest* if it meets the requirements for Strong for $\geq$ 90% of graphs, rather than $\geq$ 50%, and if for at least 95% of graphs none of the alternative distributions are favored over the power-law.

Our results are consistent with nearly all biochemical networks, at either the individual or ecosystem-level, being "super-weakly" scale-free (Fig. 2). While the power-law is better than other models, it is not itself a good model. When doing a goodness-of-fit test, we find the majority of network representations across individual and ecosystem-level networks have $p < 0.1$. This indicates there is a < 10% chance that the data is truly power-law distributed. Additionally, when compared to other distributions through log-likelihood ratios, 99% of all data sets do not favor alternative heavy tail distributions to the power law for the majority of their network-projections (Fig. 3, top row).

Where biochemical systems succeed and fail scale-free classifications. *Goodness-of-fit p-value.* The "weakest" requirement for a scale-free network introduced by Broido and Clauset stipulates at least 50% of a dataset's network-projections must have a power-law goodness-of-fit $p \geq 0.1$. For both individuals and ecosystems, only 6% of network-projections meet this requirement (Fig. 4, left column). This goodness-of-fit p-value requirement is the most restrictive of all scale-free requirements.

Tail size. Setting aside the fact each subsequent scale-free requirement builds on the requirement(s) of the preceding one, we find 98% of individual networks and 99% of ecosystem networks *do* meet the requirement of $n_{tail} \geq 50$ for a scale-free degree distribution (Fig. 4, center column).

The power-law exponent, α. Only 50% of individual-level networks and 51% of ecosystem-level networks meet the requirement that $2 < \alpha < 3$ for their degree distribution. The goodness-of-fit p value requirement, followed by the requirement constraining values of α, are the most restrictive when determining whether a biochemical network's degree distribution should be considered scale free (Fig. 4, right column).

Meeting the threshold for scale-free classification is dependent on the network representation. We find the results of each requirement listed above for classifying topology as scale-free differ across the eight network projection types for each dataset. Unsurprisingly, for most requirements, there exists a minute difference between the values observed for the largest connected component and entire graph of a given network projection type (Fig. 3, right column). Depending on the measure, there is a noticeably larger difference between the major network projection types, e.g., between bipartite, unipartite-reactions, unipartite-compounds (where all substrates participating in the same reaction are connected), and unipartite-compounds (where substrates on the same side of a reaction are not connected) (Fig. 3, right column).

Comparing to alternative distributions. Over 99% of individual and ecosystem-level datasets have 6 projections which do not favor any other distribution over the power-law (Fig. 3, top row, left column). No datasets have more than 6. The other two projections nearly always favor at least one other distribution over the power-law distribution—either the log-normal, exponential, stretched exponential, or power-law with exponential cutoff (Fig. 3, top row, right column). There are only 3 of the 6280 ecosystem-level network projections (across the 785 ecosystem-level datasets) that do not favor at least one of the alternative distributions. Oftentimes all four are favored over the power-law distribution (Fig. S1, rows 3–4). These results are identical, within 95% confidence, for both individuals and ecosystems.

Goodness-of-fit p-value. Out of all datasets, 80% of individuals and 84% of ecosystems have only a single projection type with $p \geq 0.1$ for a power-law fit to their degree distribution. This indicates the majority of datasets would still not meet the "weakest" requirement for scale-free even with a threshold that lowered the percent of a dataset's projections needed to 25% (2 networks) instead of 50% (4 networks) (Fig. 3, 2nd row, left column). The unipartite projection where substrates on the same side of a reaction are not connected (unipartite-subs_not_connected) was the most likely to satisfy $p \geq 0.1$. For the two unipartite-compound projections, the difference

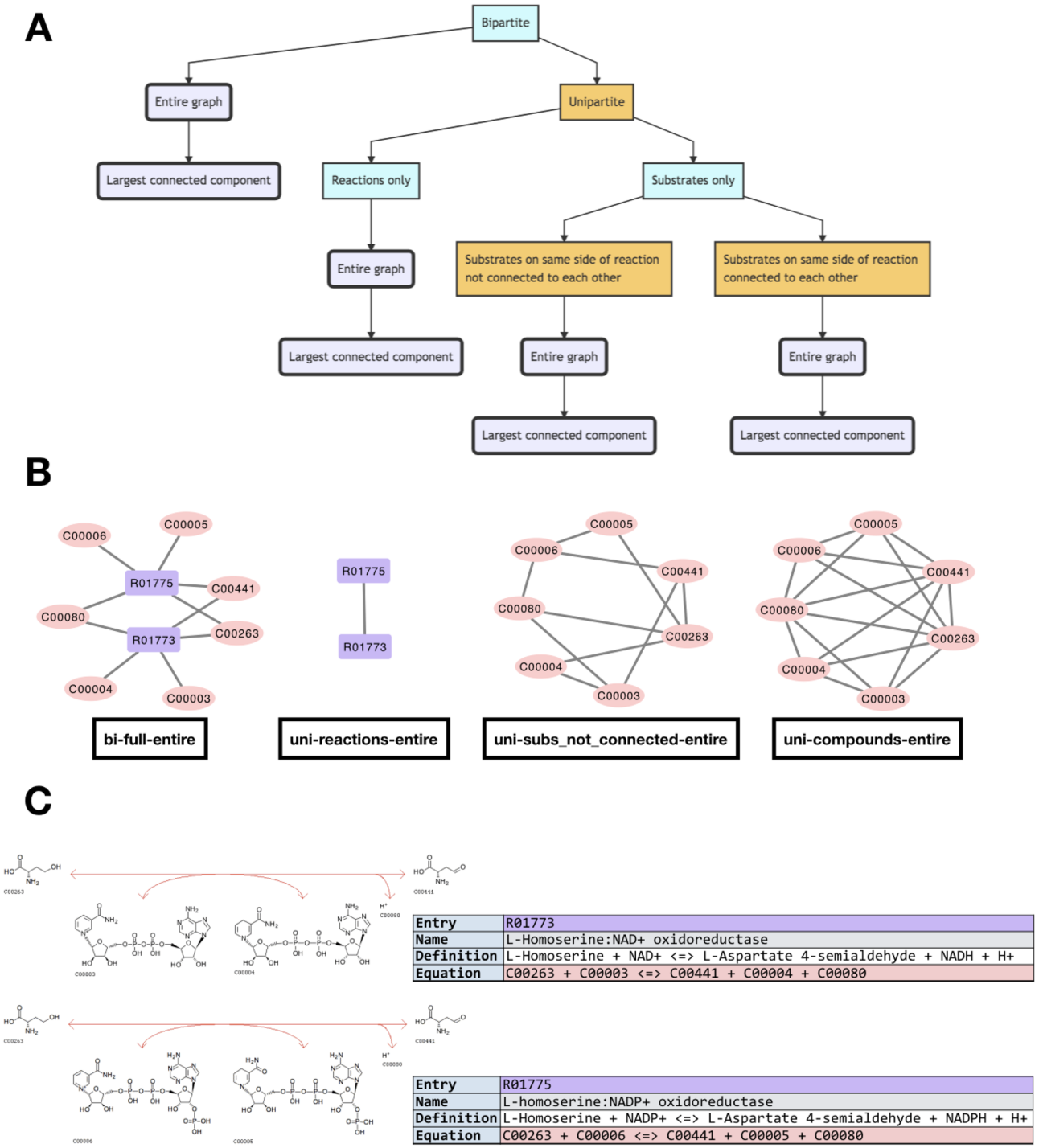

Figure 1. How biochemical datasets are decomposed into network projections. (**A**) Networks are generated from the set of reactions encoded in each genome/metagenome starting from a bipartite representation, and projecting different combinations of attributes. The bold, rounded flowchart nodes show the result of each combination of projections applied in this study. (**B**) The different network projection types of a simple example dataset, composed of two KEGG reactions: R01773 & R01775. The nomenclature used in this paper's figures is below each network visualization (in this example the entire graph is the same as the largest connected component). (**C**) How the reactions used in the network visualization example above appear in the KEGG database[46–48].

between individuals and ecosystems is within the error. The unipartite-reaction projections were the least likely to satisfy $p \geq 0.1$, which is consistent with the observation that these networks always favor an alternative distri-

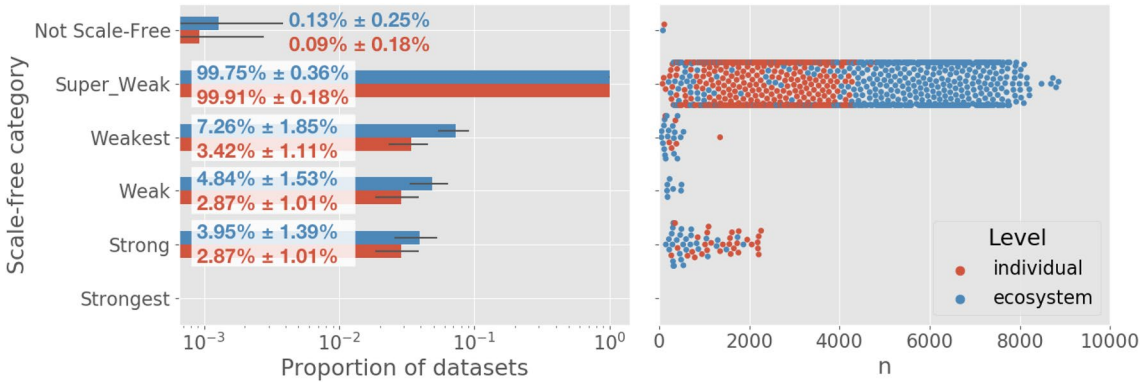

Figure 2. The vast majority of individual and ecosystem level networks are not "scale-free". *Left* Most datasets are super weak, indicating that when compared to other models, a power-law distribution is a better fit. However, the power-law distribution is not a "good" fit for most dataset network representations. No networks meet the "Strongest" criteria defined by Broido and Clauset al.[30]. Overlaid values show the percent of networks of each level which fall into each category, ±2SD. *Right* The relationship between scale-freeness and largest network size across projections (n). All datasets containing networks larger than approximately 2100 nodes have degree distributions that rule out fitting well to a power-law.

bution as a better fit to the data than the power-law (Fig. 3, 2nd row, right column). As we initially reported, the majority of datasets do not meet the p-value threshold for being considered scale-free, although ecosystems-level datasets are more likely to meet the threshold.

Tail size. Out of all datasets, 98% of individuals and 96% of ecosystems meet $n_{tail} \geq 50$ for all projection types (Fig. 3, 3rd row, left column). For 7 of the projection types, there is no difference between individuals and ecosystems, within 95% confidence (Fig. 3, 3rd row, right column).

The power-law exponent α. Out of all datasets, 95% of individuals and 97% of ecosystems meet $2 < \alpha < 3$ for 4 of 8 projection types (Fig. 3, bottom row, left column). The two types of unipartite-compound networks contribute to the datasets which meet the alpha-range requirement the majority of the time. That is, chances are if a dataset has at least 4 projection types meeting $2 < \alpha < 3$, two of them are going to be unipartite-compound network projections (Fig. 3, bottom row, right column). The results are similar for both individuals and ecosystems.

Correlation of results between projections. Because 8 different network projections are derived from a single biochemical dataset, there is reason to expect the proportions of each projection type meeting any given scale-free criteria are correlated. We therefore constructed a Pearson correlation matrix to test whether there are correlations between projections (Fig. S2). Unsurprisingly, we find that values from projections of a network's LCC and entire graph are highly correlated. All types of unipartite compound networks tend to be correlated. Values across many other projection types are barely correlated for the p-value and n_{tail} criteria. Ecosystems tend to show more correlation, across all projection types, than individuals.

Distinguishing individuals and ecosystems based on their degree distributions. *Multinomial regression.* We used multinomial regression on network and degree distribution data from the above analyses to attempt to distinguish individuals from ecosystems. Most measures cannot reliably distinguish between these two levels of organization, with only network size and network tail size data distinguishing the two levels better than chance. Using only network size, ecosystems could be correctly identified in test data 72.23% of the time, whereas individuals could be correctly identified 85.33% of the time (Fig. 5, left columns). When normalizing other measures to network size, the only one that improved in distinguishing individuals and ecosystems to be better than chance was *dexp* (Fig. S3). This is a measure of which type of distribution is favored (or neither) when doing a log-likelihood ratio test between the power-law and exponential distribution.

Random forest. Random forest classifiers are a supervised machine learning technique that use decision trees to make classifications. When using random forests to try and distinguish individuals and ecosystems based on network and distribution data, we find ecosystems can be correctly predicted 87.01% of the time, and individuals can be correctly predicted 95.82% of the time (Out of bag, OOB, error rate is 7.91%). However, the size of the network and size of the degree distribution tail once again are the best relative predictors. Without network size and tail size, the prediction accuracy drops to 79.27% for ecosystems and 94.81% for individuals (OOB error rate of 11.80%). When doing random forest classification by projection type, the prediction accuracies are still above 75% for ecosystems and 91% for individuals across all projections, which is better than multinomial regression models even when information about network size is included (Fig. 5, left columns; Table S1). Mean degree was the best predictor across all network projection types.

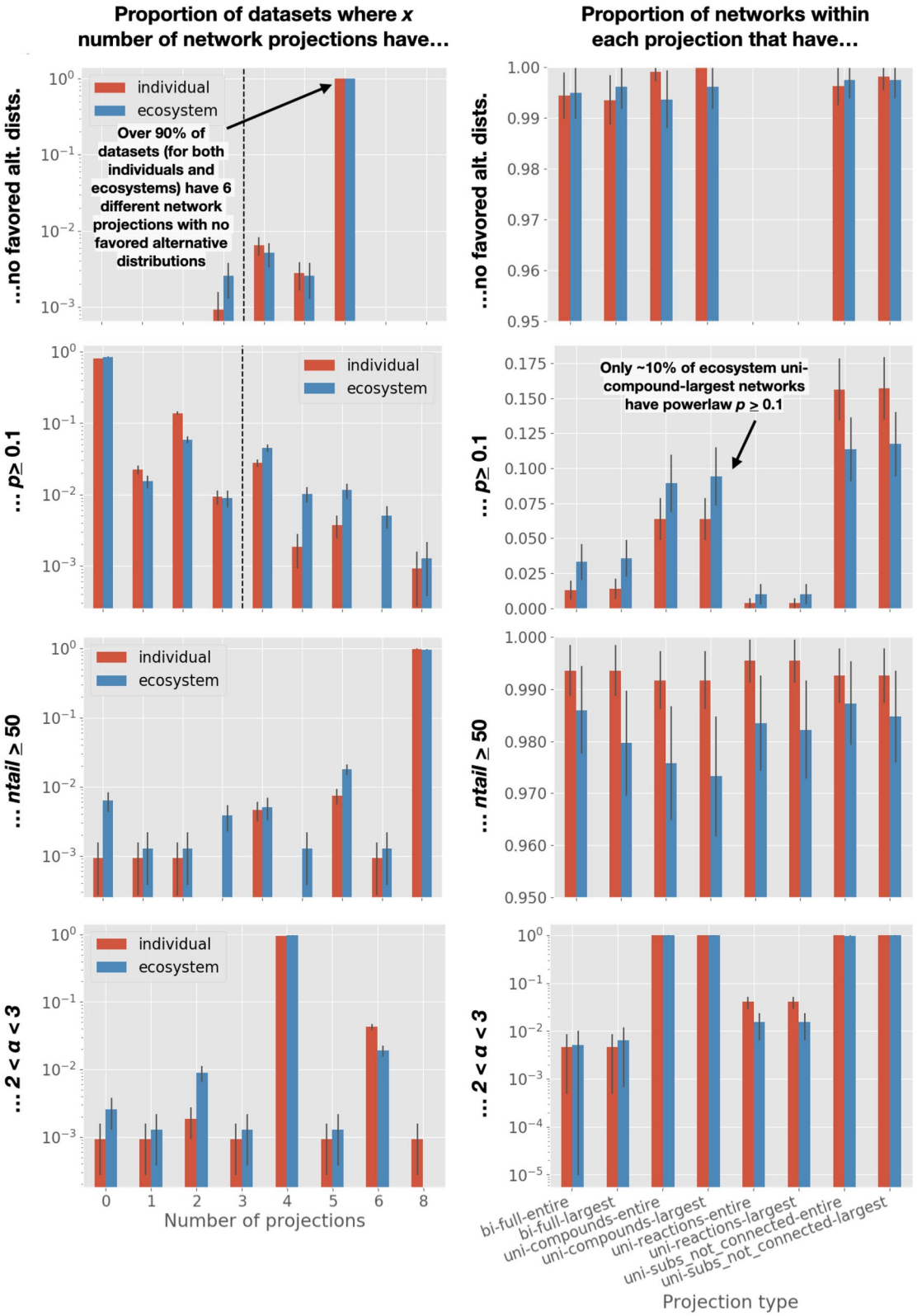

Figure 3. The number of network projections within each dataset which meet some scale-free criteria. *Left column* The number of network projections within each dataset which meet some scale free-criteria, where each dataset falls into one of nine bins. Normalized to total number of datasets in a level. Criteria from top to bottom: No alternative distributions favored over power-law in log-likelihood ratio (1st row); $p \geq 0.1$ (2nd row); $n_{tail} \geq 50$ (3rd row); $2 < \alpha < 3$ (4th row). Dashed lines show: the cutoff for number of networks in a dataset required to meet the threshold criteria for "Super-Weak" (1st row), and "Weakest" (2nd row). *Right column* The number of network projections, across all datasets, which meet some scale-free criteria, binned by projection type. Normalized to the total number of each projection within a level. Criteria same as left column. Red bars indicate individual-level datasets/networks, and blue bars indicate ecosystem-level datasets/networks. Black error bars show ±2SD.

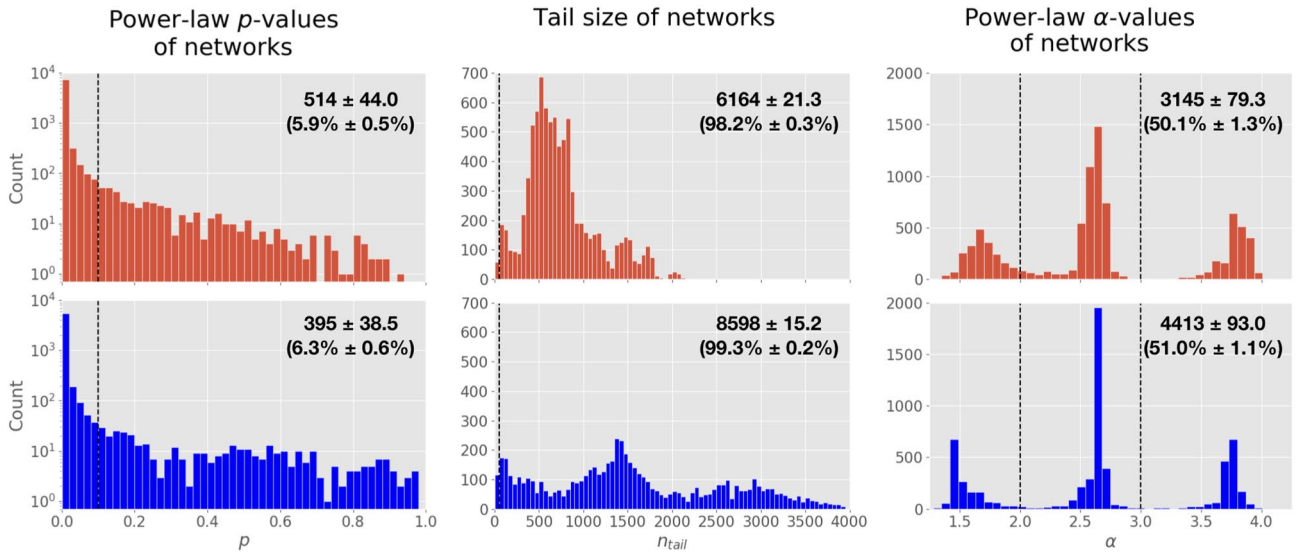

Figure 4. The distribution of p-values, tail-sizes, and power-law alpha values for biochemical network degree distributions, over all network projections. *Left column* The goodness-of-fit p-values of networks. When $p \geq 0.1$ (dashed line), it indicates that there is at least a 10% chance of the power-law distribution being a plausibly good fit to a network's degree distribution. *Center column* Tail size of networks. When $n_{tail} \geq 50$ (dashed line), it indicates that the tail of distribution is large enough to reliably fit. *Right column* Power-law exponent α values of networks. When $2 < \alpha < 3$ (between dashed lines), it indicates that a network meets the criteria of having a power-law exponent which falls into scale-free territory. The top row (in red) shows distributions for individuals. The bottom row (in blue) shows distributions for ecosystems. Insets indicate the number (and percent) of networks which meet the criteria, ± 2SD.

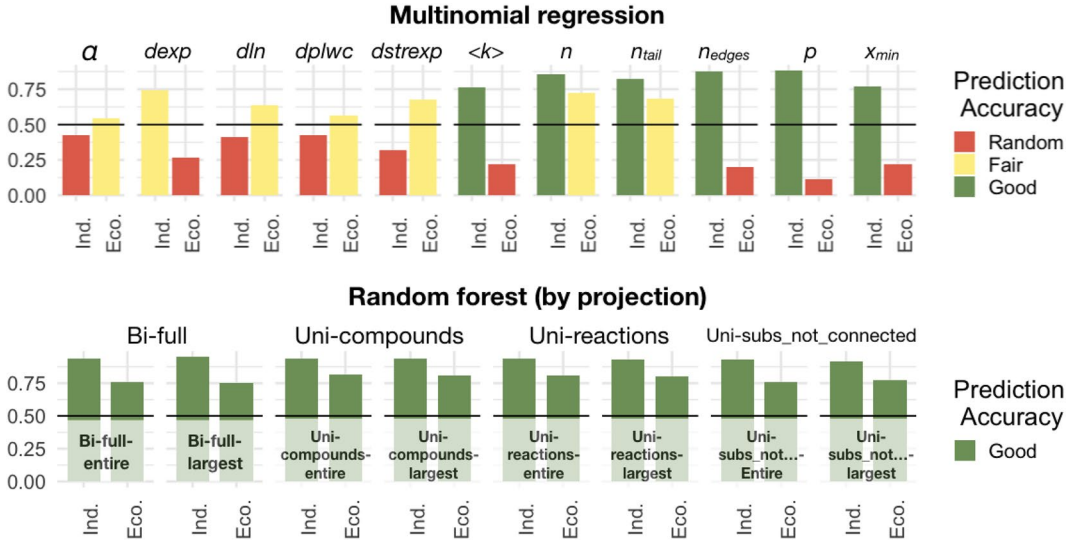

Figure 5. Predicting individuals and ecosystems from degree distribution data using multinomial regression vs. random forest. Each subplot shows the accuracy of using a particular network or statistical measure to predict whether that network data came from a biological individual or ecosystem. The top plots show prediction accuracy from using multinomial regression across all network projection types, and the bottom plots show prediction accuracy using random forest on each type of projection. The random forest classifier is much better at predicting individuals and ecosystems correctly from network data, even without direct access to network size. All random forest predictions have an accuracy of at least 75% across all projection types. Subplots measures are: power-law alpha value; log-likelihood result from power-law vs. exponential; log-likelihood result from power-law vs. log-normal; log-likelihood result from power-law vs. power-law with exponential cutoff; log-likelihood result from power-law vs. stretched exponential; the network mean degree; network node size; degree distribution tail size; network edge size; the p-value of the goodness-of-fit test for the power-law model; cutoff degree value for network tail. These are the only predictors used in the random forest classifier. Prediction accuracy is random if $\leq 50\%$, Fair if $> 50\%$, and Good if $\geq 75\%$.

Discussion

Our results indicate biochemical systems across individuals and ecosystems are, at best, only weakly scale-free. This is revealed by studying all possible projections of biochemical systems in tandem: only six of the eight network projection types analyzed favor power-law distributions over alternatives and in all cases the power-law is not itself a good fit to the data. Nonetheless, we can conclude individuals and ecosystems both share qualitatively similar degree distribution characteristics, and while this is a very coarse-grained measure of network structure, it suggests the possibility of shared principles operating across levels of organization to architect biochemical systems. The random forest distinguishability analyses demonstrate how using a combination of all the results of scale-free analyses completed in this paper can predict, better than chance, whether the data comes from individuals or ecosystems. Individuals are perhaps more tightly constrained in their network structure, based on being able to more accurately predict them based on simple network characteristics. Whether or not this structure is truly a universal property of life's chemical systems is more difficult to conclude. Based on the sample sizes, we are confident our results hold over the population of genomes and metagenomes in the JGI and PATRIC databases. However, the observed scaling is only reflective of biology universally if the databases are unbiased in sampling from all of biology on Earth, and this is impossible to know with certainty (see e.g. proposals of 'shadow life' and reports of missing biota[49,50]). Nonetheless, the fact that multiple levels and multiple projections of biochemistry reveal common structure suggests universal principles may be within reach if cast within an ensemble theory of biochemical network organization (see e.g. also[41]).

Achieving an ensemble theory for biochemistry will requires different approaches to those that have been used to apply to cases of simpler physical systems where statistics over individual components are sufficient to describe and predict their behavior. Complex systems are complex precisely because they require additional information about the structure of interactions among their many components. This challenge is well-known. However, the most effective methods for projecting these high-dimensional structures onto simple mathematical objects to enable their analysis and comparison is among the most central problems of complexity science. By contrast, in physics coarse-graining procedures are well known, but we are not so advanced in understanding complex systems that we have similarly useful tools at hand. A first challenge is to identify coarse-grained network representations, which is subject to debate. Current literature cautions against the use of unipartite graphs, as they can lead to "wrong" interpretations of some system properties, including degree[34,51]. We find instead that this conclusion is not so easy to arrive at. Whether the interpretation of a given representation is correct depends strongly on the characteristics of the degree distribution under consideration. As an example, all network projection types in our analysis, aside from unipartite reaction networks, favor power-law degree distributions over other heavy-tailed alternatives (Fig. 3, top row). For power-law α, there are a similar proportion of networks with bipartite projections and ecosystem unipartite reaction projections with $2 < \alpha < 3$ (within 2SD). However, the proportion of networks within this alpha range differ when compared using ecosystems, or any unipartite compound projections (Fig. 3, fourth row). Nearly all projections show different results for the scale-free p-value cutoff (Fig. 3, second row). While previous work[32,33] has advocated for unipartite networks (where all compounds that participate in a reaction are connected—called *uni-compounds* here), we find these overestimate the power-law goodness-of-fit p-estimates and the values for α when compared to reaction networks or bipartite networks (Fig. 3). The nuances of both similarity and difference in the structure of the same system across different projections can provide insights into the underlying system of interest, providing details that are inaccessible looking at just one projection. That is, regardless of whether or not a given projection is scale-free, all projections provide insight into the underlying system. In physics, we are accustomed to a unique coarse-grained descriptor describing all relevant features. To understand complex interacting systems, such as the systems of reactions underlying all life on Earth, it may be the case that we should forgo the allure of simple, singular models with only one coarse-grained description. Instead, to characterize living processes, it may be time to adopt and develop theory for statistical analyses over many projections in tandem.

Materials and methods

Obtaining biological data. Bacteria and Archaea data were obtained through PATRIC[52]. Starting with the 21,637 bacterial genomes available from the 2014 version of PATRIC, we created a parsed dataset by selecting one representative genome containing the largest number of annotated ECs from each genus. Unique genera (genera only represented by a single genome) were also included in our parsed data. Uncultured/candidate organisms without genera level nomenclature are left in the parsed dataset. This left us with 1152 parsed bacteria, from which we chose 361 randomly to use in this analysis. Starting with 845 archaeal genomes available from the 2014 version of PATRIC, we randomly chose 358 to use in this analysis. Enzyme Commission (EC) numbers associated with each genome were extracted from the `ec_number` column of each genome's `.pathway.tab` file.

Eukarya and Metagenome data were obtained through JGI IMG/m[53]. All 363 eukaryotic genomes available from JGI IMG/m as of Dec. 01, 2017 were used. Starting with the 5586 metagenomes available from JGI IMG/m as of June 20, 2017, 785 metagenomes were randomly chosen for this paper's analyses. Enzyme Commission (EC) numbers associated with each genome/metagenome were extracted from the list of *Protein coding genes with enzymes*, and metagenome EC numbers were obtained from the *total* category. All JGI IMG/m data used in this study were sequenced at JGI.

Because each EC number corresponds to a unique set of reactions that an enzyme catalyzes, the list of EC numbers associated with each genome and metagenome can be used to identify the reactions that are catalyzed by enzymes coded for in each genome/metagenome. We use the Kyoto Encyclopedia of Genes and Genomes (KEGG) ENZYME database to match EC numbers to reactions, and the KEGG REACTION database to identify

the substrates and products of each reaction[46–48]. This provides us with a list of all chemical reactions that a genome/metagenome's enzymes can catalyze.

Generating networks. Each genomic/metagenomic dataset is used to construct eight representations of biochemical reaction networks. We refer to each type of representation as a "network projection type" throughout the text:

1. *Bipartite graph with reaction and compound nodes*. A compound node C_i is connected to a reaction node R_i if it is involved in the reaction as a reactant or a product. Abbreviated in figures as *bi-full*.
2. *Unipartite graph with compound nodes only*. Two compound nodes C_i and C_j are connected if they are both present in the same reaction. A reaction's reactant compounds are connected to each other; a reaction's product compounds are connected to each other; and a reaction's reactant and product compounds are connected. Abbreviated in figures as *uni-compounds*.
3. *Unipartite graph with reaction nodes only*. Two reaction nodes R_i and R_j are connected if they involve a common compound. Abbreviated in figures as *uni-reactions*.
4. *Unipartite graph with compound nodes only (alternate)*. Two compound nodes C_i and C_j are connected only if they are both present on opposite sides of the same reaction. A reaction's reactant compounds are *not* connected to each other; a reaction's product compounds are *not* connected to each other; but a reaction's reactant and product compounds are connected. Abbreviated in figures as *uni-subs_not_connected*.

There exists a version of each of these four network construction methods for the largest connected component (LCC), and for the entire graph, yielding a total of eight network projections for each dataset (Fig. 1). These network projection types are signified in the figured by appending *-largest* and *-entire* to the network projection abbreviations. Some datasets may yield identical networks for their LCC and entire graph, if there is exists only a single connected component.

Assessing the power-law fit on degree distributions. As defined in Clauset[29], a quantity x obeys a power law if it is drawn from a probability distribution

$$f(x) = x^{-\alpha}, \tag{2}$$

where α, the exponent/scaling parameter of the distribution, is a constant. In order to estimate α, we follow the methods described in Clauset[29], and use an approximation of the discrete maximum likelihood estimator (MLE)

$$\hat{\alpha} \simeq 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min} - \frac{1}{2}}\right]^{-1}, \tag{3}$$

where x_{min} is the lower bound of power-law behavior in our data, and x_i, i=1,2,...,n, are the observed values x such that $x_i \geq x_{min}$. The standard error of our calculated α is given by

$$\sigma = \frac{\hat{\alpha} - 1}{\sqrt{n}} + \mathrm{O}\,(1/n), \tag{4}$$

where the higher-order correction is positive[29]. Because many quantities only obey a power-law for values greater than some x_{min}, the optimal x_{min} value must be calculated. The importance of choosing the correct value for x_{min} is discussed in detail in Clauset et al.[29]. If it is chosen too low, data points which deviate from a power-law distribution are incorporated. If it is chosen too high, the sample size decreases. Both can change the accuracy of the MLE, but it is better to err too high than too low.

In order to determine x_{min}, we use the method first proposed by Clauset et al.[54], and elaborated on in Clauset et al.[29]: we choose the value of x_{min} that makes the probability distributions of the measured data and the best-fit power-law model as similar as possible above x_{min}. The similarity between the distributions is quantified using the Kolmogorov–Smirnov or KS statistic, given by

$$D = \max_{x \geq x_{min}} |S(x) - P(x)|, \tag{5}$$

where $S(x)$ is the cumulative density function (CDF) of the data for the observations with value at least x_{min}, and $P(x)$ is the CDF for the power-law model that best fits the data in the region $x \geq x_{min}$. Our estimate of x_{min} is the one that minimizes D.

We used the github respository made available in Broido and Clauset[30] to determine the optimal x_{min} of all our degree distributions, and to subsequently calculate the MLE in order to determine the *scaling exponent* α and the *standard error* on α, σ[55].

A power-law can always be fit to data, regardless of the true distribution from which it is drawn from, so we need to determine whether the power-law fit is a good match to the data. We do this by sampling many synthetic data sets from a true power-law distribution, recording their fluctuation from power-law form, and comparing this to similar measurements on the empirical data in question. If the empirical data has similar form to the synthetic data drawn from a true-power law distribution, then the power-law fit is plausible. We use the KS statistic to measure the distance between distributions.

We use a goodness-of-fit test to generate a p-value which indicates the plausibility of a hypothesis. The p-value is defined as the fraction of the synthetic distances that are larger than the empirical distance. If p is large (close to 1), then the difference between the empirical data and the model can be attributed to statistical fluctuations alone; if it is small, the model is not a plausible fit to the data[29]. We follow the methods in Clauset et al.[29]—and implement them with the github package used in Broido and Clauset[30]—to generate synthetic datasets and measure the distance between distributions. Following these methods, we chose to generate 1000 synthetic datasets in order to optimize the trade-off between having an accurate estimation of the p-value and computational efficiency. If p is small enough ($p < 0.1$) the power law is ruled out. Put another way, it is ruled out if there is a probability of 1 in 10 or less that we would by chance get data that agree as poorly with the model as the data we have[29]. However, measuring a $p \geq 0.1$ does not guarantee that the power-law is the most likely distribution for the data. Other distributions may match equally well or better. Additionally, it is harder to rule out distributions when working with small sample sizes.

A better way to determine whether or not data is drawn from a power-law distribution is to compare its likelihood of being drawn from a power-law distribution directly to a competing distribution[29,56]. We use the exponential, stretched-exponential, log-normal, and power-law-with-cutoff distributions as four competing distributions to the power-law. While we cannot compare how the data fits between every possible distribution, comparing the power-law distribution to these four similarly shaped competing distributions helps us ensure that our results are valid.

We use the log-likelihood ratio test $\mathcal{R}$[29,56] to compare the power-law distribution to other candidate distributions,

$$\mathcal{R} = \mathcal{L}_{\mathrm{PL}} - \mathcal{L}_{\mathrm{Alt}}, \tag{6}$$

where $\mathcal{L}_{\mathrm{PL}}$ and $\mathcal{L}_{\mathrm{Alt}}$ are the log-likelihoods of the best fits for the power-law and alternative distributions, respectively. This can be rewritten as a summation over individual observations,

$$\mathcal{R} = \sum_{i}^{n_{\mathrm{tail}}} [\ell_i^{(\mathrm{PL})} - \ell_i^{(\mathrm{Alt})}], \tag{7}$$

with the log-likelihood of single observed degree values under the power-law distribution, $\ell_i^{(\mathrm{PL})}$, and alternative distribution, $\ell_i^{(\mathrm{Alt})}$, are summed over the number of model observations, n_{tail}.

If $\mathcal{R} > 0$, the power-law distribution is more likely; if $\mathcal{R} < 0$, the competing candidate distribution is more likely; if $\mathcal{R} = 0$, they are equally likely. Just like with the goodness of fit test, we need to make sure our result is statistically significant ($p < 0.01$). The methodology described here summarizes the methodology introduced by Clauset et al. (2009), and described again in Broido and Clauset[29,30] and more details such as the exact formulas for alternative distributions, and derivation of the p-value for $\mathcal{R}$ can be obtained therein.

Classifying network scaling. We classify each genomic/metagenomic dataset, as represented by the set of eight network projection types, as having some categorical degree of "scale-freeness" from "super-weak" to "strongest". This classification scheme was introduced by Broido and Clauset[30] in order to compare many networks with different degrees of complexity, and the definitions below were extracted from therein:

- *Super-Weak* For at least 50% of graphs, none of the alternative distributions are favored over the power law.

The four remaining definitions are nested, and represent increasing levels of direct evidence that the degree structure of the network data set is scale free:

- *Weakest* For at least 50% of graphs, the power-law hypothesis cannot be rejected ($p \geq 0.1$).
- *Weak* The requirements of the Weakest set, and there are at least 50 nodes in the distribution's tail ($n_{\mathrm{tail}} \geq 50$).
- *Strong* The requirements of the Weak and Super-Weak sets, and that $2 < \hat{\alpha} < 3$.
- *Strongest* The requirements of the Strong set for at least 90% of graphs, rather than 50%, and for at least 95% of graphs none of the alternative distributions are favored over the power-law.

Categorizing a network as "Super-Weak" is in effect saying that that network's degree distribution data is *better* modeled by a power-law fit than alternative distributions. This is independent of whether or not the power-law model is a *good* fit to the data, which is what is what the "Weakest" and "Weak" definitions emphasize. A network may be classified as "Super-Weak" without meeting any of the nested definition's criteria. Similarly, a network may be classified as "Weak" without meeting the criteria in the "Super-Weak" definition. We believe this framework is a proper way to classify the degree-distributions of biochemical networks, given that there are many different accepted ways to represent biochemical reactions as networks, and each has their pros and cons[32–34].

Standard error and correlation. The black error bars on each plot represent 2 standard deviation (2SD) around the sample proportion $\hat{p}$ (the height of the bar, which we also refer to as the mean). This is equivalent to 2 standard error around the mean (2SEM), or a 95% confidence interval for the true population proportion p (true population mean). Standard deviation was calculated by treating each category as a binomial distribution, meaning the standard deviation is given by:

$$\sqrt{\frac{p(1-p)}{n}}. \tag{8}$$

Although the errors for each plot's categories are calculated independently, there is co-variance between many of them. This is especially true for the right column of Fig. 3, where all bars of a color total to a fixed number of datasets, with each dataset falling into one of the 8 network projection type bins. Because of this, we also calculated the correlations between each network projection type, across both individuals and ecosystems (Fig. S2). The correlation matrices were calculated by using the pandas function `DataFrame.correlation(method='pearson')` on a matrix of binomially distributed True/False values representing whether each dataset passed or failed specific scale-free criteria for p-value, tail size, or power-law exponent value (α), for each network-projection.

Classifying levels of biology using degree distribution data. We used two different statistical methods, multinomial regression and random forest classifiers, in conjunction with the scale-free classification scheme above in order to test if individuals and ecosystems were distinguishable based on their degree distribution characteristics.

Multinomial regression. For our multinomial regression, the response class is the biological level (individual or ecosytem), and a single network or statistical measure is the dependent variable. In order to control for over fitting the training data was composed of an equal number of samples from each level. The number of networks used for training data was chosen to be equal in size to 80% of all ecosystem projections, because there were less ecosystem datasets used than individual datasets. This corresponded to 80% of 6280 networks (of all projection types), or 5024 networks. The model was tested on the 20% of the data that it was not trained on. This process was repeated 100 times and the average model error is reported in the results and Fig. 5, left columns. The `multinom` and `predict` functions from the R-package `nnet` were used to do the multinomial regression.

Random forest classifiers. We used a random forest to attempt to classify networks as falling into the category of individuals or ecosystems. In the first scenario, we used 11 predictors: power-law alpha value (α); log-likelihood result from power-law vs. exponential (*dexp*); log-likelihood result from power-law vs. log-normal (*dln*); log-likelihood result from power-law vs. power-law with exponential cutoff (*dplwc*); log-likelihood result from power-law vs. stretched exponential (*dstrexp*); the network mean degree ($< k >$); network node size (n); degree distribution tail size (n_{tail}); network edge size (n_{edges}); the p-value of the goodness-of-fit test for the power-law model (p); and cutoff degree value for network tail (x_{min}). In the second scenario, we repeated the random forest without the three predictors which can be directly used to quantify the size of a network (n, n_{tail}, and n_{edges}). In the third scenario, we repeated the random forest without the three predictors on each network projection type independently. For each scenario, we randomly split our data in two halves: one for training, and one for testing (for the third scenario, each training and testing set is 1/8 as large as for the first two scenarios, since we run the classifier on each network projection type independently). In all scenarios, we use the `randomForest` function from the R-package `randomForest` for classification. Three features were used to construct each tree (`mtry=3`), which is $\approx \sqrt{n_{features}}$, with 100 trees generated each time (enough time for the out-of-bag, or OOB, estimate of the error rate to level off).

Received: 29 September 2020; Accepted: 19 February 2021
Published online: 22 March 2021

References

1. Strogatz, S. H. Exploring complex networks. *Nature* **410**(6825), 268 (2001).
2. Albert, R. & Barabási, A. L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**(1), 47 (2002).
3. Dorogovtsev, S. N. & Mendes, J. F. Evolution of networks. *Adv. Phys.* **51**(4), 1079–1187 (2002).
4. Barabasi, A. L. & Oltvai, Z. N. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **5**(2), 101 (2004).
5. Newman, M., Barabasi, A. L. & Watts, D. J. *The Structure and Dynamics of Networks* Vol. 19 (Princeton University Press, 2011).
6. Barabási, A. L. *et al. Network Science* (Cambridge University Press, 2016).
7. Newman, M. *Networks* (Oxford University Press, 2018).
8. Girvan, M. & Newman, M. E. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**(12), 7821–7826 (2002).
9. Milo, R. *et al.* Network motifs: Simple building blocks of complex networks. *Science* **298**(5594), 824–827 (2002).
10. Newman, M. E. The structure and function of complex networks. *SIAM Rev.* **45**(2), 167–256 (2003).
11. Albert, R. Scale-free networks in cell biology. *J. Cell Sci.* **118**(21), 4947–4957 (2005).
12. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D. U. Complex networks: Structure and dynamics. *Phys. Rep.* **424**(4–5), 175–308 (2006).
13. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **407**(6804), 651–654 (2000).
14. Koonin, E. V., Wolf, Y. I. & Karev, G. P. *Power Laws, Scale-Free Networks and Genome Biology* (Springer, 2006).
15. Guimera, R. & Amaral, L. A. N. Functional cartography of complex metabolic networks. *Nature* **433**(7028), 895 (2005).
16. Tanaka, R. Scale-rich metabolic networks. *Phys. Rev. Lett.* **94**(16), 168101 (2005).
17. Gosak, M. *et al.* Network science of biological systems at different scales: A review. *Phys. Life Rev.* **24**, 118 (2018).
18. Whitesides, G. M. Is the focus on "molecules" obsolete?. *Annu. Rev. Anal. Chem.* **6**, 1–29 (2013).
19. Cronin, L. & Walker, S. I. Beyond prebiotic chemistry. *Science* **352**(6290), 1174–1175 (2016).

20. Walker, S. I. & Mathis, C. Network theory in prebiotic evolution. In *Prebiotic Chemistry and Chemical Evolution of Nucleic Acid* (ed. Menor-Salvan, C.) 263–291 (Springer, 2018).
21. Barabási, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**(5439), 509–512 (1999).
22. Albert, R., Jeong, H. & Barabási, A. L. Internet: Diameter of the world-wide web. *Nature* **401**(6749), 130 (1999).
23. Carlson, J. M. & Doyle, J. Highly optimized tolerance: A mechanism for power laws in designed systems. *Phys. Rev. E* **60**(2), 1412 (1999).
24. Albert, R., Jeong, H. & Barabási, A. L. Error and attack tolerance of complex networks. *Nature* **406**(6794), 378–382 (2000).
25. Barabási, A. L. Scale-free networks: A decade and beyond. *Science* **325**(5939), 412–413 (2009).
26. Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. *Internet Math.* **1**(2), 226–251 (2004).
27. Barabási, A. L. & Bonabeau, E. Scale-free networks. *Sci. Am.* **288**(5), 60–69 (2003).
28. Li, L., Alderson, D., Doyle, J. C. & Willinger, W. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Math.* **2**(4), 431–523 (2005).
29. Clauset, A., Shalizi, C. R. & Newman, M. E. Power-law distributions in empirical data. *SIAM Rev.* **51**(4), 661–703 (2009).
30. Broido, A. D. & Clauset, A. Scale-free networks are rare. *Nat. Commun.* **10**(1), 1–10 (2019).
31. Khanin, R. & Wit, E. How scale-free are biological networks. *J. Comput. Biol.* **13**(3), 810–818 (2006).
32. Holme, P. Model validation of simple-graph representations of metabolism. *J. R. Soc. Interface* **6**(40), 1027–1034 (2009).
33. Holme, P. & Huss, M. Substance graphs are optimal simple-graph representations of metabolism. *Chin. Sci. Bull.* **55**(27–28), 3161–3168 (2010).
34. Montanez, R., Medina, M. A., Sole, R. V. & Rodríguez-Caso, C. When metabolism meets topology: Reconciling metabolite and reaction networks. *Bioessays* **32**(3), 246–256 (2010).
35. Klarreich, E. Scant evidence of power laws found in real-world networks. In *Quanta Magazine*, 15 (2018).
36. Cohen, R., Erez, K., Ben-Avraham, D. & Havlin, S. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.* **85**(21), 4626 (2000).
37. Bollobás, B. & Riordan, O. Robustness and vulnerability of scale-free random graphs. *Internet Math.* **1**(1), 1–35 (2004).
38. Pastor-Satorras, R. & Vespignani, A. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**(14), 3200 (2001).
39. Zhou, B., Meng, X. & Stanley, H. E. Power-law distribution of degree-degree distance: A better representation of the scale-free property of complex networks. *Proc. Natl. Acad. Sci.* **117**(26), 14812–14818 (2020).
40. Smith, E. & Morowitz, H. J. *The Origin and Nature of Life on Earth: The Emergence of the Fourth Geosphere* (Cambridge University Press, 2016).
41. Kim, H., Smith, H. B., Mathis, C., Raymond, J. & Walker, S. I. Universal scaling across biochemical networks on Earth. *Sci. Adv.* **5**(1), eaau0149 (2019).
42. Featherstone, D. E. & Broadie, K. Wrestling with pleiotropy: Genomic and topological analysis of the yeast gene expression network. *Bioessays* **24**(3), 267–274 (2002).
43. Guelzim, N., Bottani, S., Bourgine, P. & Képès, F. Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.* **31**(1), 60 (2002).
44. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540 (2004).
45. Kaiser, M. A tutorial in connectome analysis: Topological and spatial features of brain networks. *Neuroimage* **57**(3), 892–907 (2011).
46. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**(1), 27–30 (2000).
47. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**(D1), D457–D462 (2015).
48. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**(D1), D353–D361 (2016).
49. Davies, P. C. *et al.* Signatures of a shadow biosphere. *Astrobiology* **9**(2), 241–249 (2009).
50. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci.* **113**(21), 5970–5975 (2016).
51. Klamt, S., Haus, U. U. & Theis, F. Hypergraphs and cellular networks. *PLoS Comput. Biol.* **5**(5), e1000385 (2009).
52. Wattam, A. R. *et al.* Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.* **45**(D1), D535–D542 (2016).
53. Markowitz, V. M. *et al.* IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* **40**(D1), D115–D122 (2011).
54. Clauset, A., Young, M. & Gleditsch, K. S. On the frequency of severe terrorist events. *J. Conflict Resolut.* **51**(1), 58–87 (2007).
55. Broido, A.D. SFAnalysis (2017). https://github.com/adbroido/SFAnalysis.
56. Alstott, J., Bullmore, E. & Plenz, D. Powerlaw: A Python package for analysis of heavy-tailed distributions. *PLoS ONE* **9**(1), e85777 (2014).

Acknowledgements

We thank the Emergence@ASU team (especially Doug Moore, Cole Mathis, and Jake Hanson) for feedback through various stages of this work.

Author contributions

H.B.S., H.K. and S.W. conceived of the idea. H.B.S. performed the analysis. H.B.S., H.K. and S.W. wrote the manuscript.

Funding

The funding was provided by National Aeronautics and Space Administration (NNX15AL24G S02).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-85903-1.

Correspondence and requests for materials should be addressed to S.I.W.

Reprints and permissions information is available at www.nature.com/reprints.