Supplementary Materials for

Universal Scaling in Biochemical Networks

Hyunju Kim, Harrison B. Smith, Cole Mathis, Jason Raymond and Sara I. Walker

correspondence to: sara.i.walker@asu.edu

Materials and Methods

Terminology

The label “Synthetic” in text, table, or figure refers to what is referenced in the main text as “random genome datasets” or “random genome networks”.

The label “Random” in text, table, or figure refers to what is referenced in the main text as “random reaction datasets” or “random reaction networks”.

References to “archaea parsed” (or any variation thereof) refers to a subset of all analyzed archaea genomes. Starting with all archaea genomes, we selected 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. The “bacteria parsed” dataset was created in the same way.

Rationale for choosing data to be analyzed

We calculate all network measures on the largest connected component (LCC) of each network, for the following reasons: 1. Several network measures only make sense to calculate on connected components (e.g. diameter, average shortest path), focusing on the LCC therefore permits all network measures implemented in our study to be calculated for all networks; 2. The largest connected components have the vast majority of nodes (>90%) for the vast majority of networks in each dataset (the only exception is the random reaction networks, of which only ~76% have a largest connected component with at least 90% of a network’s nodes). **See Table S1 and Fig. S1 for distribution of sizes of the LCC by dataset.**

Comparing candidate degree distributions

In our analyses, the goodness of a distribution’s fit to a particular dataset (e.g., a network’s degree distribution) is assessed comparative to other distributions. Practically, this means fits from two candidate distributions are compared using a log-likelihood ratio (R) using standard methods (*19*, *28*). Here we report the normalized log-likelihood ratio, $R/(n * \sqrt{\_})$, where *n* is the number of data points ≥ x_{min}, and **σ** is the standard deviation on R), along with its significance (p). If the p value is small (p < 0.1), then the observed sign of R is likely a reliable indicator of which candidate distribution is a better fit to the dataset. For our analyses, a positive R indicates a preference for the power law distribution (we always put R in the numerator of the log-likelihood ratio). For each dataset, we compared a power law distribution to three other common distributions for goodness of fit: exponential, log-normal, and truncated power law. This produces an R and p for each comparison.

We conclude *it is not unreasonable to fit the dataset to a power law distribution* if both of the following conditions are met:
a) When compared to an exponential distribution, R>0 and p<0 (i.e. the power law distribution is favored compared to the exponential distribution, and thus the distribution is heavy-tailed(*20*, *29*).
b) When compared to each other distribution, R>0 or p≥0.1 (i.e. the power law distribution is favored or the observed sign of R is not statistically significant)

[*Note*: In the main text, we report on a version of plausibility where we compared the power law distribution to only two other distributions: exponential and lognormal; thus excluding the truncated power law in the main discussion.]

We are less concerned with verifying if our data truly follows a power law distribution than verifying that a power law distribution is a reasonable description to use—even if other descriptions may also be reasonable. We therefore did not perform a goodness-of-fit test to determine, in absolute terms, whether the data is drawn from a power-law distribution (*23*). For example, even if we used bootstrapping and the Kolmogorov-Smirnov test to determine the validity of a power law fit to the datasets irrespective of any other distributions, we may find that our data does not fit strictly to a power law distribution. We use the power law distribution because it is frequently implemented to describe heterogeneity in biochemical network structure, and importantly it can be described with a single scaling parameter, making comparisons across different datasets easier than in the case of comparing log-normal distributions which have two scaling parameters. Additionally, we are interested in knowing if our dataset is more favorably fit to a heavy-tail distribution than to a non-heavy-tail distribution (e.g., the exponential distribution).

We use the Python *power law* package to find x_{min} (see **Fig. S4**) as well as to calculate the log-likelihood ratio and its significance for each candidate distribution comparison (*20*). For more details and nuances of distribution fitting and comparison, see Clauset et al., 2009 (*19*).

Power law Fits for Degree Distributions

Using our above definition of power law plausibility, the vast majority of all networks could plausibly fit a power law distribution, when compared to the candidate distributions of exponential, lognormal, and truncated power law distributions (**Table S2 col 2**). If we remove the comparison to truncated power law distributions, that number increases (**Table S2 col 3**).

Additionally, we compare the α values of networks which can plausibly be fit by a power law distribution. α references the scaling exponent of the power law distribution, given by $P(k) = ck^{-\alpha}$. The distribution of α values obtained across all datasets can be seen in **Fig. S5**.

We do this across all networks for two scenarios: 1. When we include the truncated power law as a candidate distribution for comparison, and 2. When we remove the truncated power law as a candidate distribution for comparison. The results indicate not comparing to a truncated power law causes the spread of α to increase, while the range of values and median α value remain similar. The results can be seen for real and random genome/random reaction networks in **Fig. S2**. The only exception to this result is for the random reaction network dataset, where roughly half of the networks fit significantly better to a truncated power law distribution than a power law distribution, when a truncated power law distribution is included in the comparison. In case where we do not compare to the truncated power law, the distribution is noticeably shifted downward in addition to the increased spread of the data. We highlight the difference between these two groups of data by separately plotting the subsets of the random reaction networks which are and are not power law plausible as a boxplot and scatterplot in **Fig S3 a and b**, respectively.

Fitting network measure scaling and permutation tests

For each network measure, a scaling relationship was fit as a function of the size of the largest connected component (LCC) of the network. For each measure, three different models were tested, a power law of the form $y = y_0 x^{\beta}$, a linear relationship of the form $y = \beta x + y_0$, and a quadratic function of the form $y = \beta_1 x + \beta_2 x^2 + y_0$, for both the assortativity measures, the preferred fit was also compared to a constant $y = \beta$. The preferred model was chosen as the one which minimized cross validation errors, according to 10-fold cross validation, across the entire data set.

Once a model was chosen, a simulated permutation test was performed to determine whether the scaling relationship for a given attribute was the same for ecosystems and individuals or if it was distinct (*24*). We took as the null hypothesis the scaling relationship across different levels of organization are constant, and used the fitted scaling parameters (for individuals and ecosystems) as the test statistic. We used fitted 1,000,000 resamples of the complete dataset to estimate the likelihood of the fit for individuals (or ecosystems) to have been drawn randomly from the complete dataset. We performed this test for both the ecosystem and individuals, if there was a difference in the estimated likelihoods we took the greater of the two. These likelihoods are the (two-sided) p values reported in Table 1 (main text). The same procedure was followed to determine the distinguishability of ecosystem networks with the randomized controls (random genome networks, and random reaction networks).

To estimate the true scaling parameters, and 95% confidence intervals a bootstrap sample of 100,000 was used for each network attribute(*24*). If the permutation test allowed us to reject the hypothesis of a constant scaling relationship across individuals and ecosystems to a confidence greater than 0.01, the scaling parameters were estimated separately for the individuals and ecosystems, otherwise the complete dataset was fit. The scaling parameters (and confidence intervals) for distinct domains were also estimated using a bootstrap of 100,000 samples.

For scaling fits and confidence intervals see **Table S5**.

Predicting evolutionary domain from topology

To demonstrate topological features of genomes from different domains are distinct, multinomial regression was used(*30*). Specifically, we implemented models where the domain of the network was the response class and a single topological feature, normalized by the size of the largest connected component (LCC) of the network was the dependent variable. We found topological features of networks alone were often not predictive of the domain but the ratio of the topological properties to the size of the network was often more predictive. Prior to the regression these normalized topological measures were scaled and centered(*24*). The regression was implemented in base R using the **glm(..),** function. 10-fold cross validation was used to estimate the prediction accuracy of any one measure, which is reported in Table 1 (main text).

Obtaining genomic and metagenomic information

Genomes (PATRIC)

Archaea and bacteria genomic datasets were obtained from PATRIC(10). Enzyme commision (EC) numbers were obtained from "ec_number" column in the pathway data of each taxon. Eukarya genomic datasets were obtained from the Joint Genome Institute's (JGI) integrated microbial genomes database and comparative analysis system (IMG/M)(11). All eukarya data used in this study was sequenced at JGI. All EC numbers used to construct eukarya biochemical networks were obtained from the list of total enzymes associated with each eukaryote. EC numbers were used in conjunction with KEGG enzyme and reaction data in order to build biochemical networks for each taxon.

Metagenomes (JGI)

Metagenomic data was obtained from JGI IMG/M(11). All metagenomic data used in this study was sequenced at JGI. All EC numbers used to construct metagenomic biochemical networks were obtained from the list of total enzymes associated with each metagenome.

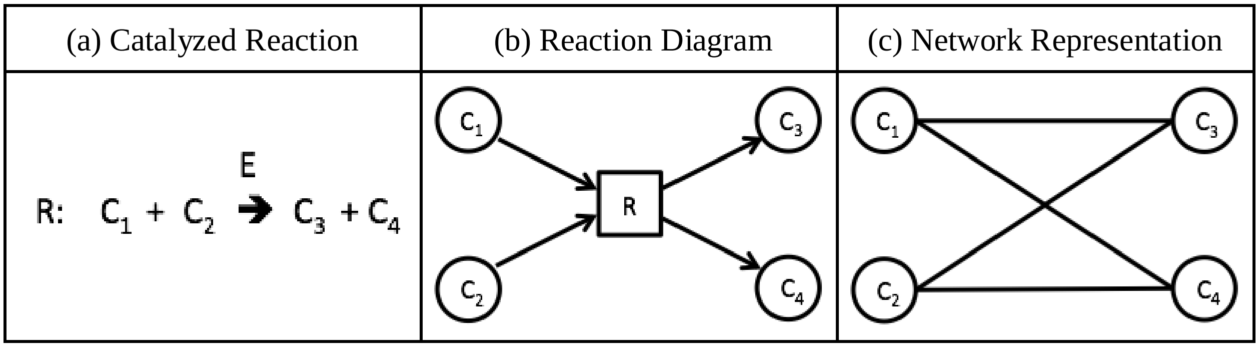
These EC numbers were used in conjunction with KEGG enzyme and reaction data in order to build biochemical networks for each metagenome.

Enzyme Classes (EC numbers)

In the main text, what we refer to as enzyme classes are EC numbers. All reaction data associated with each EC number was obtained from KEGG(*12*).

Constructing networks

In this study, we construct three different types of biochemical reaction networks: biological networks, random genome networks and random reaction networks. These biochemical reaction networks consist of chemical compounds that are involved in biochemical reactions: two chemical compounds are connected to each other when one is a reactant and the other is a product of the same biochemical reaction. The process to encode a biochemical reaction as the network representation can be described with the diagram below as follows:



(a) Suppose that a chemical reaction R catalyzed by an enzyme E is given, which transforms chemical compounds C1 and C2 to C3 and C4.

(b) The reaction, R, can be described in a reaction diagram, or a directed bipartite network representation, where the reactants C1 and C2 are connected to the reaction node and the products C3 and C4 are connected as products from the same reaction.

(c) The network representation of the reaction R shows how the reaction information is embedded in the network. In the network representation, nodes are substrates and a reactant is connected directly to a product if they are connected to the same reaction in the corresponding reaction diagram.

Regardless of the types of networks, all chemical network representations in this paper follow the same methods. Therefore, the distinctions between different types of biochemical reaction networks come from how we select reactions to be included in each network, which is described below. Note that all edges in the networks in this paper are represented as undirected and unweighted since our interests lie on the presence or absence of particular reactions in given networks and, in principle, all biochemical reactions can happen in both directions depending on the environment.

Biological Networks

For each biological network, we include all catalyzed biochemical reactions annotated in each genome or metagenome. More specifically, we consider three different levels of organization: individual organisms, ecosystems and the biosphere. For the construction of individual networks, we utilize the genome data of 21,637 bacterial taxa and 845 archaeal taxa from the Pathosystems Resource Integration Center (PATRIC)(*10*), as well as 77 eukaryotic taxa from the Joint Genome Institute (JGI)(*11*). From this data, we obtain the set of classes of enzymes for each genome. All reactions catalyzed by this set of enzymes and present in the Kyoto Encyclopedia of Genes and Genomes (KEGG)(*12*) database are included in the network representation of the corresponding genome. Similarly, for the network representation of each of the 5587 ecosystems from JGI, we include all reactions catalyzed by the ecosystem's coded enzymes, provided they are catalogued in the KEGG dataset. Finally, for the biosphere network, we include all catalyzed reactions in KEGG.

Random Genome Networks

To construct a random genome network, we sample genome level networks uniformly at random from the set of all individual organisms in our data set and merge them into one random genome network. When a set of multiple individual networks are merged, every node and edge present in any individual network are added to the resulting network with equal weight regardless of how many individual networks include them.

We built random genome networks with individual networks sampled from only archaea, only bacteria, only eukarya and from integration of all the three domains. The number of individual networks randomly selected and merged into a random genome network is defined as its rank. For this study, we generated 10 random genome networks with ranks 1 - 200 for archaea, 1 - 200 for bacteria, 1 - 77 for eukarya, 1 - 447 for all three domains. The maximum rank for each domain is determined so that it is larger than minimum number of distinctive individual networks that are needed to be integrated to contain all reactions annotated in the corresponding domain. To find the maximum ranks, we investigate how the number of distinct reactions increases as individual networks are integrated one by one in the order of decreasing size of the networks for each domain and for all three domain together as shown in **Fig. S8**.

In total, we generated 2,000 random genome networks from 730 individual archaea networks, 2,000 from 21,213 individual bacteria networks, 770 from 77 individual eukarya networks, and 4770 from merging all individual networks in the three domains.

Random Reaction Networks

In this paper, random reaction networks are generated by merging randomly sampled reactions from all chemical reactions from KEGG data regardless of whether a known enzyme is cataloged for the reaction. We note 31.46% of chemical compounds in the biosphere network are not included in the genomic data in our study, therefore our construction uniformly sampling the entire KEGG database, the random reaction networks can include enzymatically catalyzed reactions not included in our genomic data. Nonetheless our sampling procedure is biased to generate networks with similar biochemistry to that of the genomic networks, due to reasons explained in the main text (compounds common to all three domains tend to be highly connected (participate in many reactions) such that a uniform sampling procedure yields random networks biased to include the most common compounds used by life). As shown in **Fig. S9**, most biological networks for real individual organisms and ecosystems contain 200 - 5000 reactions. Hence, to build similar size of random reaction networks to real individual organisms and ecosystems, we selected the total number of reactions in each network from the range between 200 and 5000, sampling for each size the appropriate number of reactions from KEGG data uniformly and at random. Merging these into networks, we constructed 5,000 random reaction networks in total.

Topological Measures

To characterize the topology of biochemical networks, we utilized some of the most frequently used topological measures. These measures are well established and detailed descriptions can be found in(*18*, *31*, *32*)). Below, we briefly review these measures and related terms. For computing each measure, we used functions provided by Python software package, NetworkX(*33*).

The topological measures implemented in this paper include average degree, average clustering coefficient, average shortest path length, assortativity (degree-degree correlation), attribute assortativity (assortativity pattern with the number of reactions as a node attribute), and node and edge betweenness as well as the total number of reactions. Note we applied these measures on the largest connected component of every networks in this study. When every pair of nodes in a subset of a given network are connected through series of edges in the network, the subset is called connected component and the largest subset amongst such subsets is called the largest connected component or the giant component.

The average degree, degree sequence and degree distribution: The degree of a node i is the total number of connections between i and rest of the network and is denoted as k_i. The average degree $< k >$ is average of $k_i$for all nodes in a given network. The degree distribution is the sequence of degrees for all nodes in a given network and usually it is ordered in a non-decreasing order (rank ordered). The degree distribution is the probability distribution over values of degrees, which provides the probability for a randomly selected node to have particular degree.

The average clustering coefficient: The local clustering coefficient for a node, i, measures the local density of edges in a network by considering the number of connected pairs of neighbor nodes of i. Hence, the clustering coefficient for a node i is defined as,

$$C_i = \frac{2L_i}{k_i(k_i-1)}$$

where k_i is the degree of node i and L_i is the number of edges between the k_i neighbors of node i. The large values of C_i indicates the highly interconnected neighborhood of i. We computed the average of C_i over all nodes in each network to measure the degree of clustering of the entire network. C_i is measured by using a Networkx method **clustering(..)**.

The average shortest path length: The shortest path length between a given pair of two nodes is defined as the minimum number of edges connecting the two nodes in a given network. The shortest path length between two nodes is measured by using a Networkx **shortest_path_length(..).** We averaged the shortest path length for every pair of nodes in a given network over the entire network.

The attribute assortativity (The assortativity for the number of local reactions): Assortativity measures the tendency of two nodes with similar properties to be connected in a given network. The assortativity coefficient proposed by Newman(*34*, *35*) is formulated as follows:

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b}$$

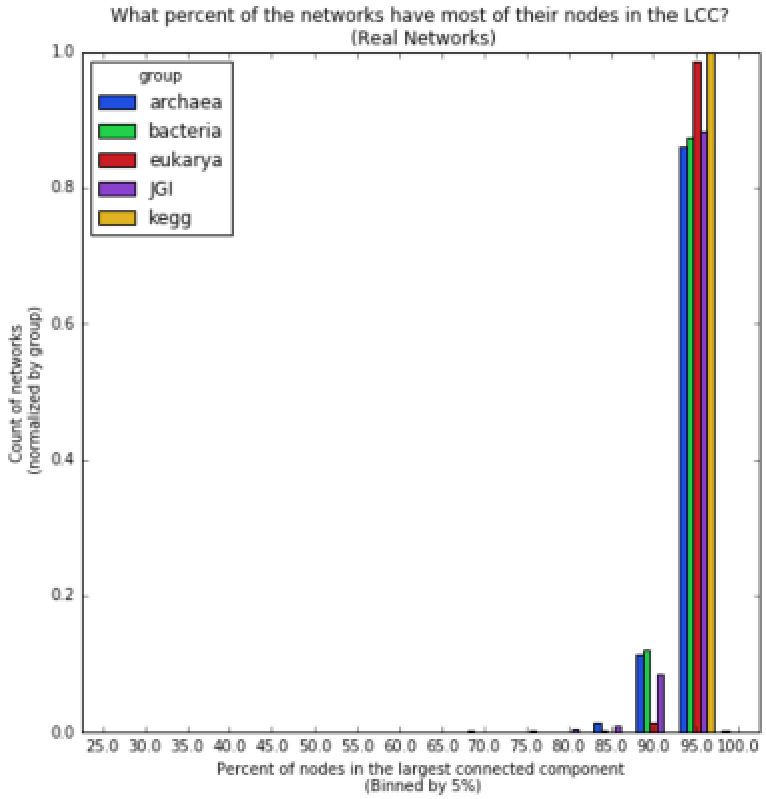
where $e_{xy}$is defined as the fraction of edges between a node with value x and one with value y for a given node attribute, and a_x and b_y are the fraction of edges coming into and going out from nodes of value x and y respectively. $\_{a}$ and $\_{b}$ are the standard deviations of the distributions of a_x and b_y. Hence, on undirected network in our study, $a_x = b_y$ and $\_{a}\_{b} = \_^{2}$. For our study, we consider this assortativity coefficient when the node attribute is the number of local reactions, i.e. reactions for each node to be involved. Also, it is referred to as *attribute assortativity* in this paper. For any network, $-1 \ r \ 1$. If $r < 0$, the network is assortative, meaning nodes in the network tend to be connected to other nodes with similar number of local reactions. If $r < 0$, nodes in the network tend to be paired to other nodes with different number of local reactions. We used a Networkx method **attribute_assortativity_coefficient(..)** to measure the attribute assortativity for our study.

The assortativity (the degree correlation coefficient): When the considered characteristics of nodes is degree, the assortativity coefficient becomes the degree correlation coefficient. It is called *assortativity* in this paper. Similar to the attribute assortativity above, If $r < 0$, the network is assortative, i.e. nodes with similar degree tend to be connected to each other. If $r > 0$, the network is disassortative, i.e. nodes in it tend to be paired to other nodes with different degrees. For any network, $-1 \quad r \quad 1$. It is measured by using a Networkx method **degree_assortativity_coefficient(..)**.

The node (edge) betweenness: Betweenness centrality of a node, $B(v)$ is defined as (*36–38*),

$$B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

where _$(s,t)$and _$(s,t|v)$denote the number of all shortest paths from s to t, and the number of the shortest paths through a given node v, respectively. Replacing _$(s,t|v)$with _$(s,t|e)$ for an edge, one can also formulate the edge betweenness. Under the assumption importance of connections is equally distributed amongst all shortest paths between each pair of nodes, the node (edge) betweenness can be considered as a measure of degree of influence of the given node (edge) over connectivity of the given network. We computed the betweenness for each node and edge by implementing Networkx methods **betweenness_centrality(..)** and **edge_betweenness_centrality(..)**, respectively.



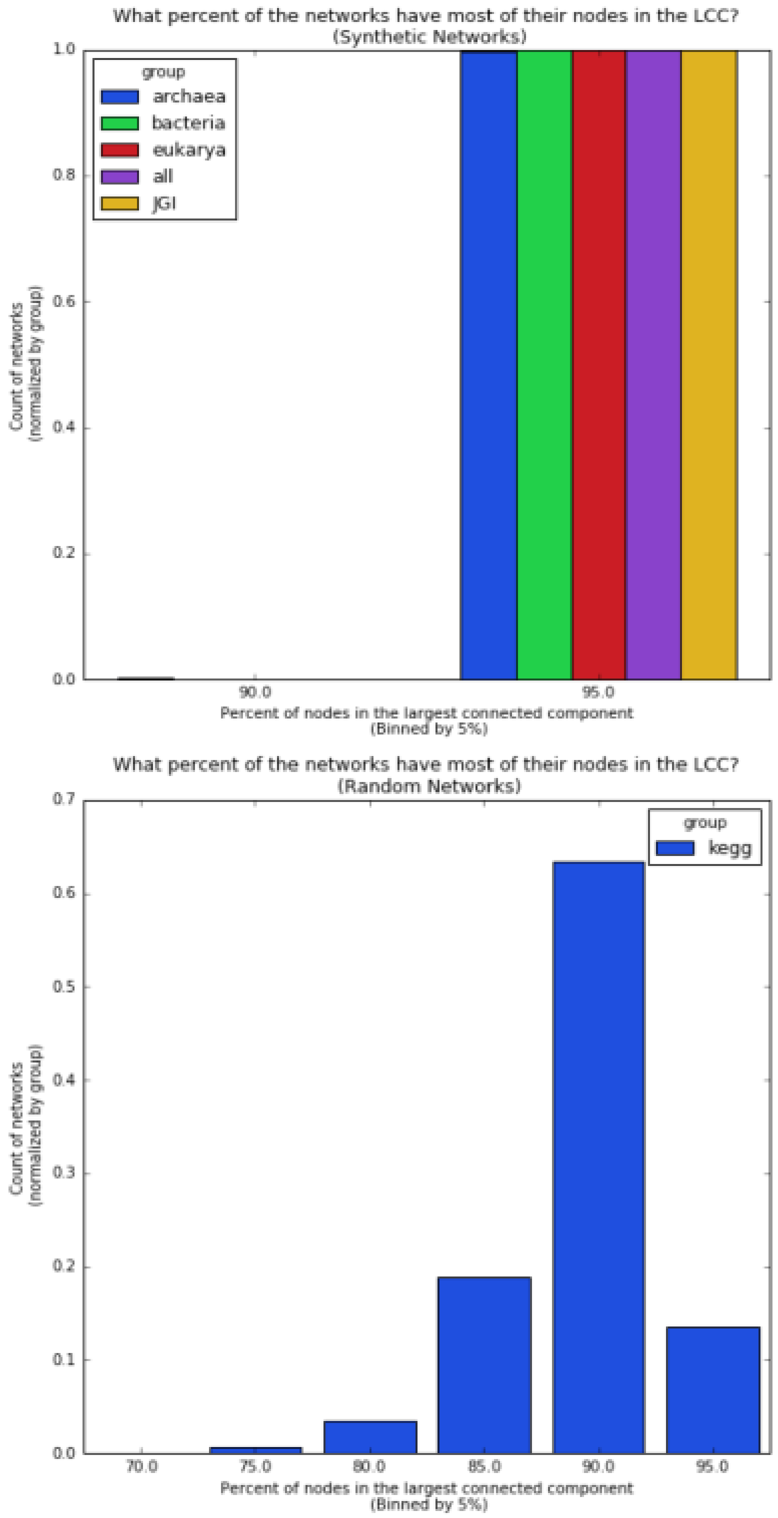
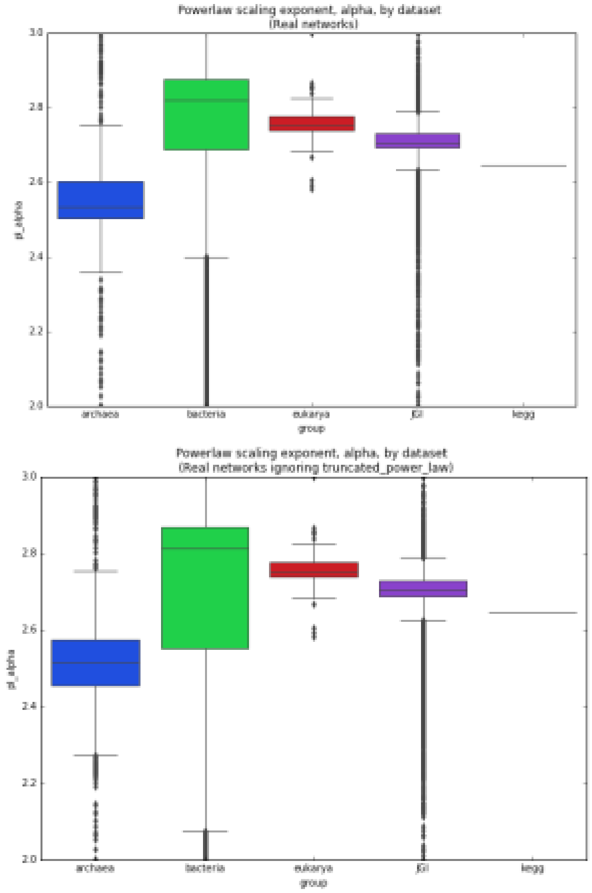


Fig. S1. The percentage of networks in each dataset with at least the percentage of nodes in the largest connected component (minimums specified on x-axis). From top to bottom: All real

datasets, all random genome datasets (labeled as synthetic), all random reaction datasets (labeled as random).



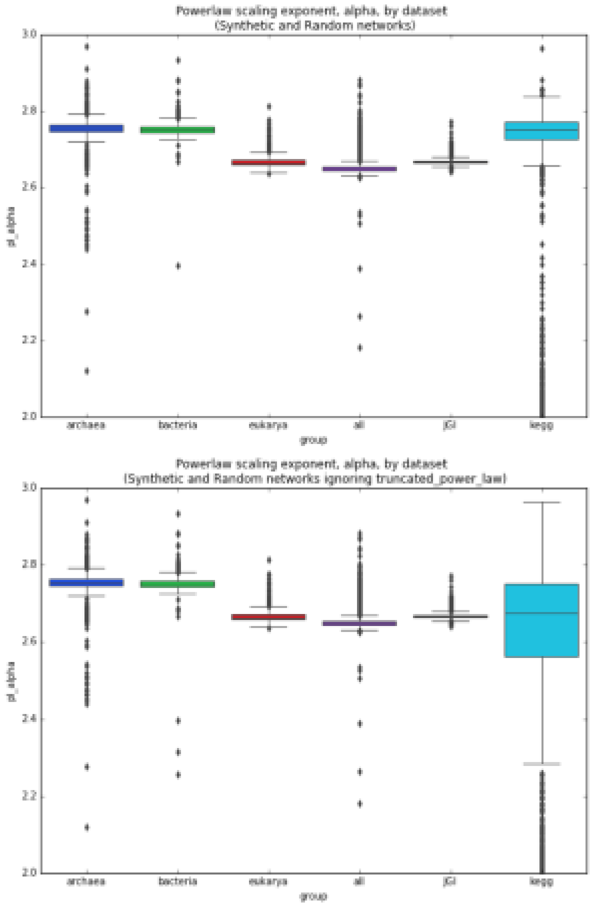


Fig. S2. The spread in alpha values for networks which can be plausibly fit to a power law distribution. The top row includes only real networks, and the bottom row includes only random genome (labeled synthetic) and random reaction networks (labeled random). The left column is for candidate distribution comparisons, which includes comparing to a truncated power law. The right column is for candidate distribution comparisons where the truncated power law is not compared to the power law. In the bottom row, the label 'kegg' refers to the random reaction networks.

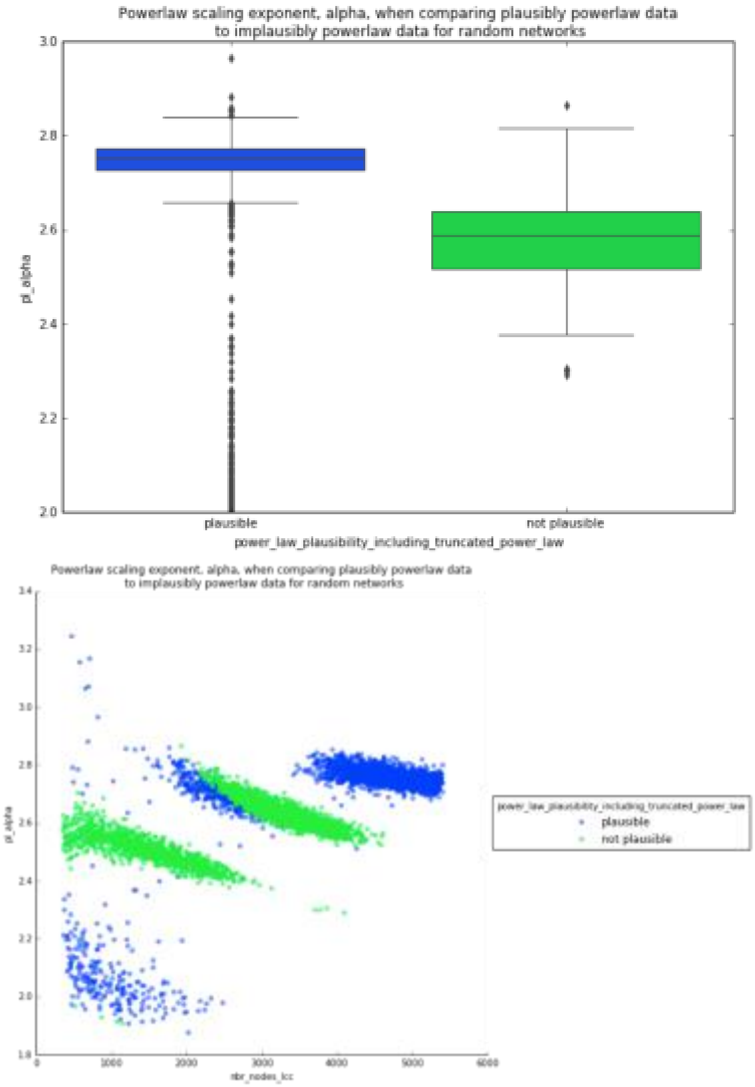
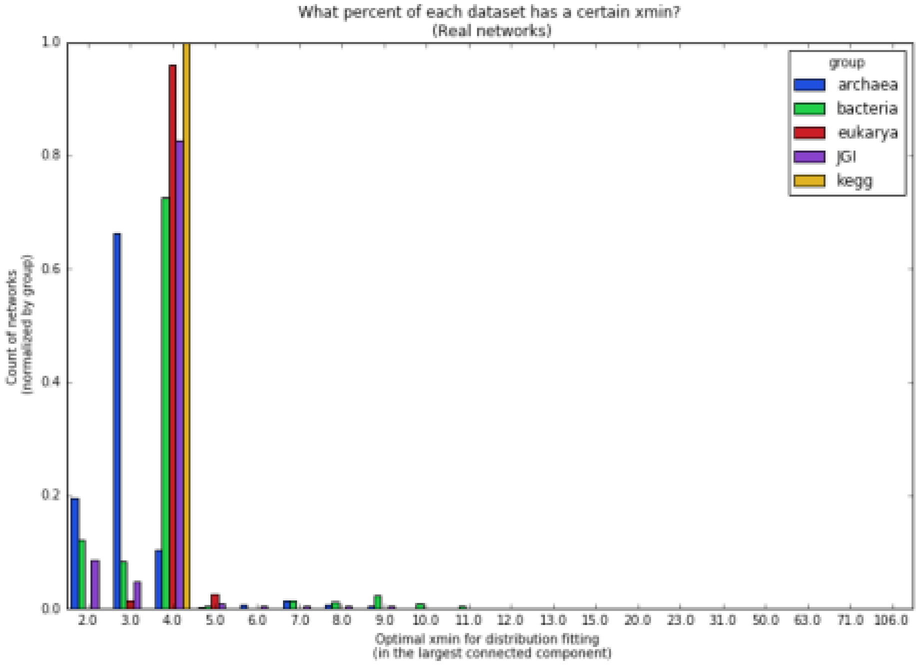
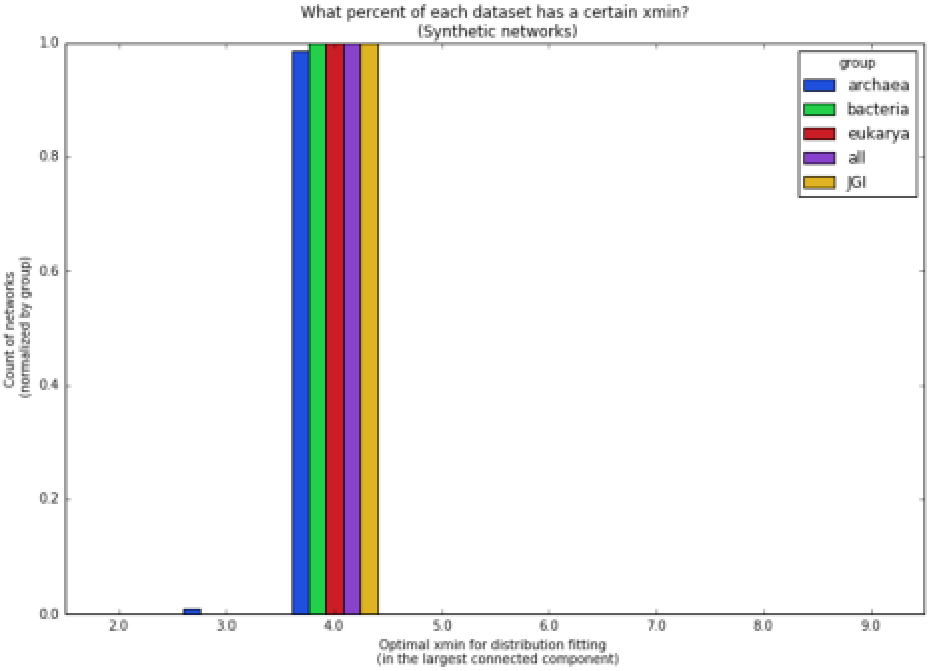


Fig. S3. Top: Boxplot showing the alpha value of the subset of random reaction networks which can plausibly be fit to a power law (blue) and the subset of networks which cannot be plausibly fit to a power law (green), when including comparisons to the truncated power law distribution. Bottom: Scatterplot showing the alpha value of the subset of random reaction networks which can plausibly be fit to a power law (blue) and the subset of networks which cannot be plausibly fit to a power law (green), when including comparisons to the truncated power law distribution. The x-axis shows the size of the network of each scatter point. The networks in green would be a plausible

fit to a power law distribution if we did not compare to the closely related truncated power law distribution.


What percent of each dataset has a certain xmin?
(Real networks)
group
archaea
bacteria
eukarya
JGI
kegg
Count of networks
(normalized by group)
0.0
0.2
0.4
0.6
0.8
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
15.0
20.0
23.0
31.0
50.0
63.0
71.0
106.0
Optimal xmin for distribution fitting
(in the largest connected component)


What percent of each dataset has a certain xmin?
(Synthetic networks)
group
archaea
bacteria
eukarya
all
JGI
Count of networks
(normalized by group)
0.0
0.2
0.4
0.6
0.8
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
Optimal xmin for distribution fitting
(in the largest connected component)

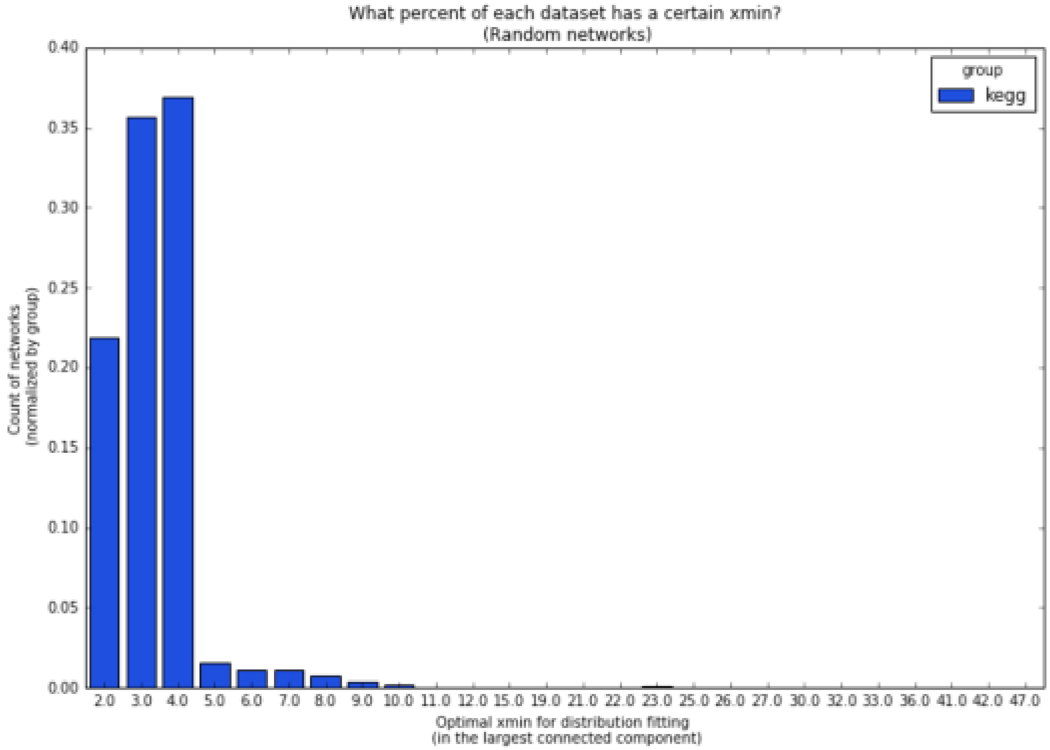


Fig. S4. The percent of each dataset with a given x_{min}. From top to bottom: real networks, random genome networks (titled synthetic networks), and random reaction networks (titled random networks).

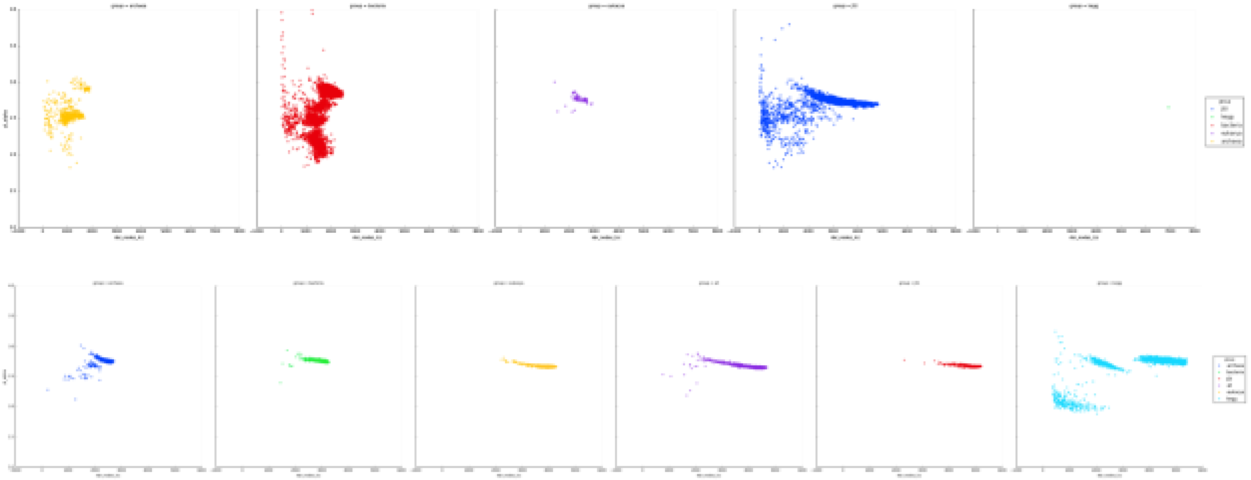


Fig S5. Scatterplot of alpha as a function of network size. Top: Biological networks. Bottom: Random genome and random reaction networks (labeled as KEGG).

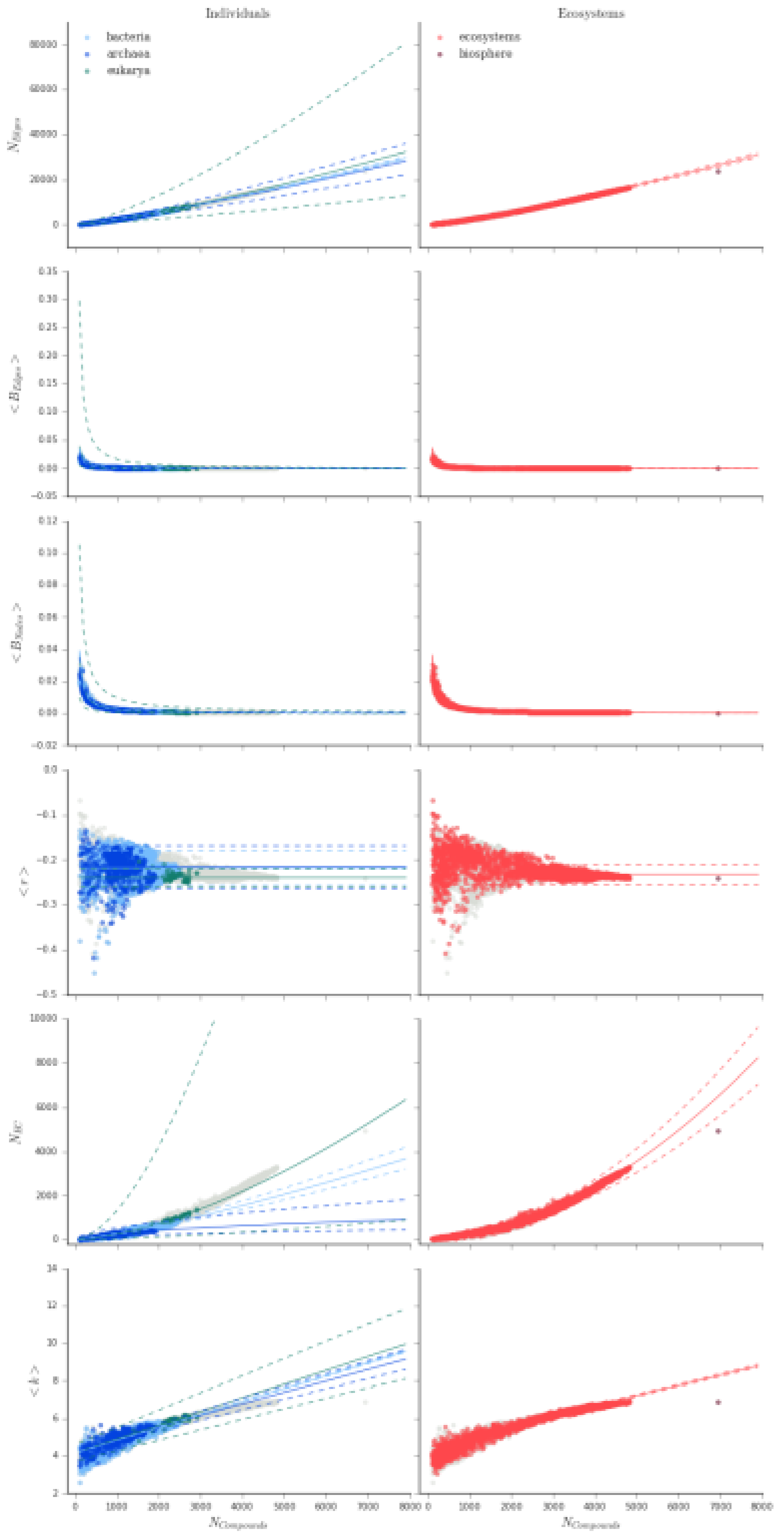


Fig. S6. Scaling behavior for additional network topology measures to those shown in Fig. 1. From top to bottom, number of edges (N_{Edges}), average edge betweenness ($<N_{Edge}>$), average node betweenness ($<N_{Node}>$), average assortativity ($<r>$), number of enzyme classes (N_{EC}).

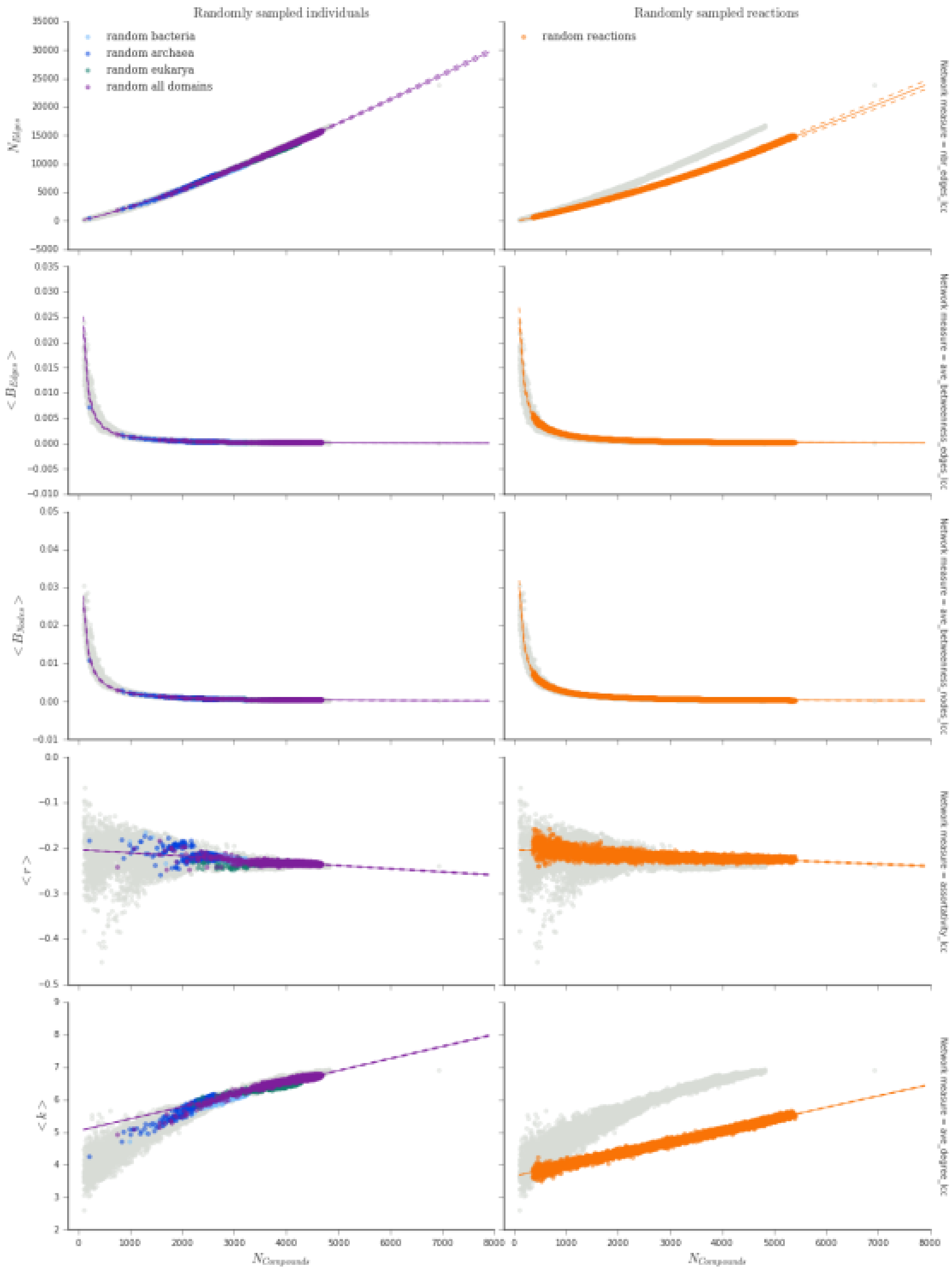


Fig. S7. Scaling behavior for additional network topology measures to those shown in Fig. 3. From top to bottom: number of edges (N_{Edges}), average edge betweenness (<N_{Edge}>), average node betweenness (<N_{Node}>), average assortativity (<r>).

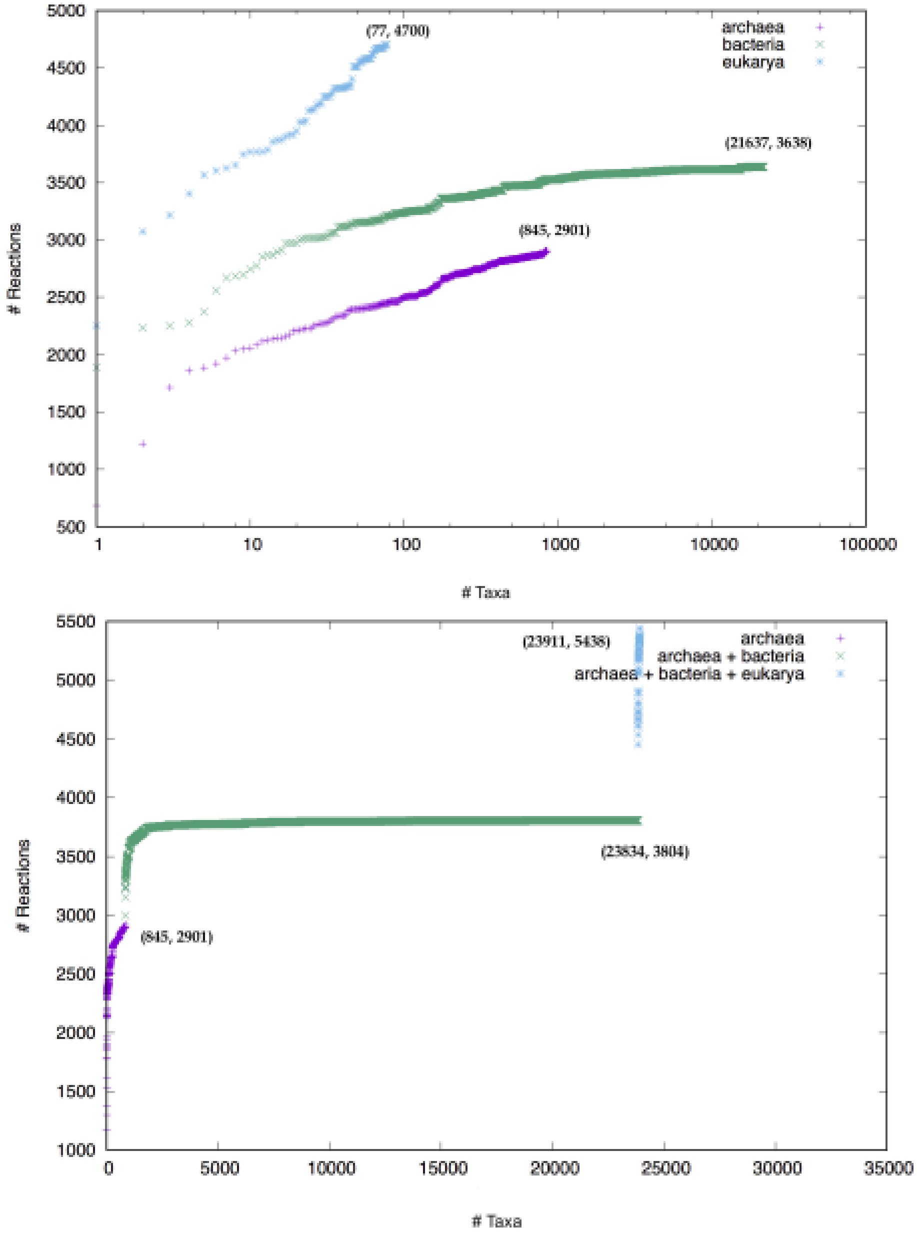


Fig. S8. The increase in the number of distinct reactions when individual organism networks are merged starting from the genome-level network with the largest number of reactions for each domain (Top) and for all the three domain combined (Bottom). Top: Shown are the trajectories for merging archaea (purple), bacteria (green) and eukarya (blue). Bottom: Shown are trajectories for successively merging archaea (purple), bacteria (green) (merged to the complete archaea network), and eukarya (blue) (merged to the complete archaea and bacteria network).

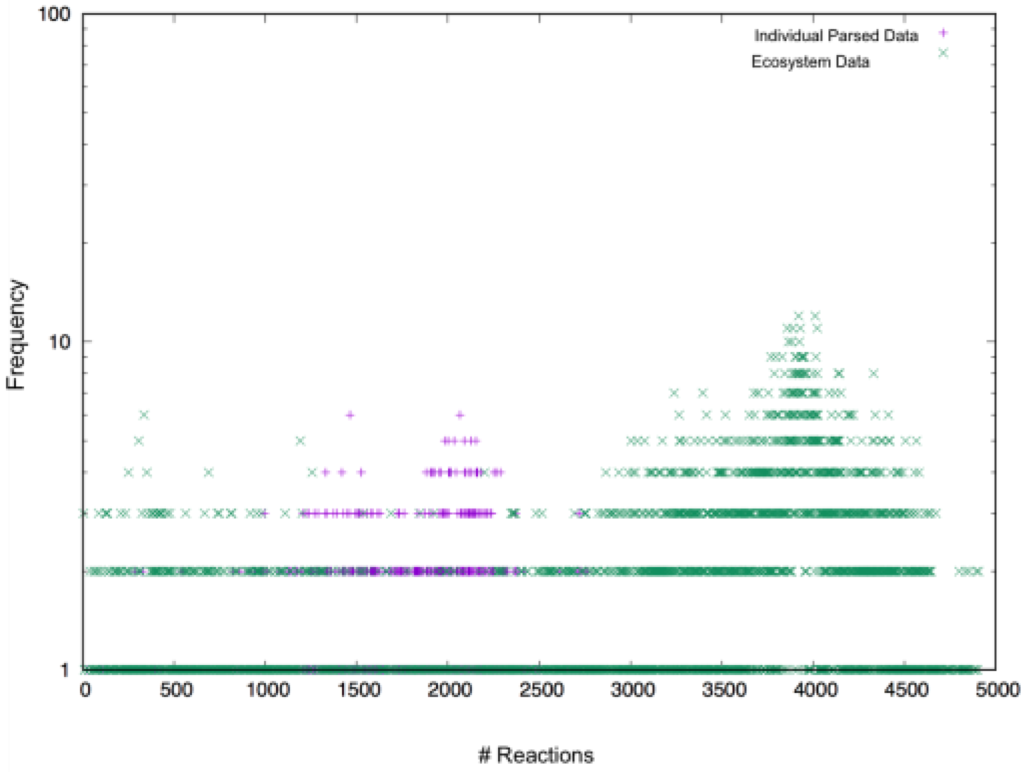


Fig. S9. The frequency of biochemical networks with a given number of reactions in the individual parsed (See Terminology) networks (purple) and ecosystem networks (green).

Table S1: Percentage of networks in each dataset with x% of nodes in the LCC

	group	>85%	>90%	>95%
Real	Arcahaea	99.17	97.75	86.39
	Arch_pars	98.49	97.49	80.90
	Bacteria	99.84	99.65	87.53
	Bact_pars	100.00	100.00	75.63
	Eukarya	100.00	100.00	98.70
	JGI	98.10	97.06	88.42
	KEGG	100.00	100.00	100.00
Random genome	Arcahaea	100.00	100.00	99.75
	Arch_pars	100.00	100.00	99.17
	Bacteria	100.00	100.00	100.00
	Bact_pars	100.00	100.00	99.56
	Eukarya	100.00	100.00	100.00
	All	100.00	100.00	100.00
	JGI	100.00	100.00	100.00
Random reaction	KEGG	95.72	76.86	13.54

Table S2: Percentage of networks plausibly fit to a power law.

	groups	# nets	% plausible (w/truncated)	% plausible (w/o truncated)
Real	Arcahaea	844	81.87	98.82
	Arch_pars	198	73.23	98.48
	Bacteria	21631	87.96	98.43
	Bact_pars	1153	90.81	98.61
	Eukarya	77	100.00	100.00
	JGI	5545	92.37	98.90
	KEGG	1	100.00	100.00
Random genome	Arcahaea	2000	99.90	99.90
	Arch_pars	1200	99.5	99.83
	Bacteria	2000	99.90	100.00
	Bact_pars	1800	100.00	100.00
	Eukarya	770	100.00	100.00
	All	4770	100.00	100.00
	JGI	4000	100.00	100.00
Random reaction	KEGG	5000	48.22	96.90

Table S3: Percentage of networks with preferred degree distributions fit functions which are heavy tailed.

	groups	# nets	% heavy tail preferred (exponential fit is significant)	% heavy tail ambiguous (no fit significant over another)
Real	Arcahaea	844	0.00	1.20
	Arch_pars	198	0.00	2.53
	Bacteria	21631	0.00	0.23
	Bact_pars	1153	0.00	0.00
	Eukarya	77	0.00	0.00
	JGI	5545	0.00	1.12
	KEGG	1	0.00	0.00
Random genome	Arcahaea	2000	0.00	0.05
	Arch_pars	1200	0.00	0.00
	Bacteria	2000	0.00	0.00
	Bact_pars	1800	0.00	0.00
	Eukarya	770	0.00	0.00
	All	4770	0.00	0.00
	JGI	4000	0.00	0.00
Random reaction	KEGG	5000	0.00	1.16

Table S4: Optimal x_{min} for candidate distribution fitting, shown are percentage of networks in each dataset with x_{min} at or below x.

Dataset	Name	2 or 3	<=4
Real	Arcahaea	85.80	96.09
	Arch_pars	83.92	95.48
	Bacteria	20.33	93.00
	Bact_pars	19.95	84.30
	Eukarya	1.30	97.40
	JGI	13.27	96.06
	KEGG	0.00	100.00
Random genome	Arcahaea	1.05	99.65
	Arch_pars	2.08	99.25
	Bacteria	0.15	100.00
	Bact_pars	0.06	99.72
	Eukarya	0.00	100.00
	All	0.06	99.94
	JGI	0.00	100.00
Random reaction	KEGG	57.52	94.46

Table S5: Scaling parameters for topological measures with 95% confidence intervals. See SI text for description of the column parameters.

beta	betaM	betaP	level	scaling	y.var	alpha	alphaP	alphaM	group
-0.2307793119	-0.2534135103	-0.2081451134	ecosystem	mean	assortativity_lcc	0	0	0	JGI
-0.2164502257	-0.255753717	-0.1771467343	individual	mean	assortativity_lcc	0	0	0	all
-0.2150657196	-0.2617221709	-0.1684092684	individual	mean	assortativity_lcc	0	0	0	archaea
-0.2164004508	-0.2554764538	-0.1773244477	individual	mean	assortativity_lcc	0	0	0	bacteria
-0.2385421647	-0.2581214935	-0.2189628359	individual	mean	assortativity_lcc	0	0	0	eukarya
-0.2034697694	-0.2040738983	-0.2028707001	syn_individual	linear	assortativity_lcc	-6.96E-06	-6.82E-06	-7.11E-06	all
-0.2033751815	-0.204134781	-0.2026202076	ranRxn_individual	linear	assortativity_lcc	-4.53E-06	-4.34E-06	-4.72E-06	ranRxn_individual
-0.002237782086	-0.01242700197	0.007951437799	ecosystem	mean	attribute_assortativity_lcc	0	0	0	JGI
0.004635407956	-0.01862052503	0.02789134095	individual	mean	attribute_assortativity_lcc	0	0	0	all
0.01114971242	-0.01764965817	0.039949083	individual	mean	attribute_assortativity_lcc	0	0	0	archaea
0.004493296291	-0.01856432585	0.02755091843	individual	mean	attribute_assortativity_lcc	0	0	0	bacteria
0.001780111	-0.007092809337	0.01065303134	individual	mean	attribute_assortativity_lcc	0	0	0	eukarya
-0.07577713199	-0.07646216882	-0.07508969488	ranRxn_individual	linear	attribute_assortativity_lcc	1.16E-05	1.18E-05	1.14E-05	ranRxn_individual
0.0009791817545	0.0005687775301	0.001410427812	syn_individual	linear	attribute_assortativity_lcc	-9.13E-07	-8.16E-07	-1.02E-06	all
2.681983848	2.596582868	2.771908934	ecosystem	powerlaw	ave_betweenness_edges_lcc	-1.333557743	-1.323097687	-1.344452509	JGI
2.63010053	2.592990457	2.668421148	individual	powerlaw	ave_betweenness_edges_lcc	-1.324922687	-1.320155573	-1.329794933	all
1.489701578	1.285518794	1.687031638	individual	powerlaw	ave_betweenness_edges_lcc	-1.171415133	-1.143838195	-1.198802682	archaea
2.78753471	2.752213466	2.822319685	individual	powerlaw	ave_betweenness_edges_lcc	-1.345838956	-1.341168967	-1.35048308	bacteria
3.18240934	2.138639068	4.635360669	individual	powerlaw	ave_betweenness_edges_lcc	-1.400705784	-1.269771971	-1.58570036	eukarya
1.599689335	1.55262854	1.649563527	ranRxn_individual	powerlaw	ave_betweenness_edges_lcc	-1.149566377	-1.143467349	-1.15610303	ranRxn_individual
1.912863709	1.17307362	2.37572118	syn_ecosystem	powerlaw	ave_betweenness_edges_lcc	-1.240466967	-1.153490235	-1.294853881	all
1.966673527	1.918606421	2.015112105	syn_individual	powerlaw	ave_betweenness_edges_lcc	-1.243668436	-1.237628536	-1.249670977	all
1.851532151	1.804906112	1.897822017	ecosystem	powerlaw	ave_betweenness_nodes_lcc	-1.136804548	-1.131139768	-1.142582387	JGI
2.012405653	1.988850845	2.036306455	individual	powerlaw	ave_betweenness_nodes_lcc	-1.158145161	-1.154871499	-1.161399378	all
1.390358005	1.263802468	1.518275936	individual	powerlaw	ave_betweenness_nodes_lcc	-1.076803844	-1.058399699	-1.094883419	archaea
2.133002115	2.110282827	2.155462494	individual	powerlaw	ave_betweenness_nodes_lcc	-1.174130262	-1.171122796	-1.177169135	bacteria
1.975193201	1.014326179	2.515695368	individual	powerlaw	ave_betweenness_nodes_lcc	-1.154817191	-1.03556797	-1.224547758	eukarya
1.372048855	1.335198467	1.410103253	ranRxn_individual	powerlaw	ave_betweenness_nodes_lcc	-1.059532228	-1.054624502	-1.064610887	ranRxn_individual

0.9577048055	0.6015624276	1.210896266	syn_ecosystem	powerlaw	ave_betweenness_nodes_lcc	-1.028028494	-0.9859516223	-1.057494185	all
1.237871182	1.20011364	1.273240277	syn_individual	powerlaw	ave_betweenness_nodes_lcc	-1.059686319	-1.055122584	-1.064194637	all
0.1133435466	0.1113121285	0.1154322204	ecosystem	linear	ave_clustering_coeff_lcc	3.32E-05	3.37E-05	3.26E-05	JGI
0.09072950006	0.08992450513	0.0915470737	individual	linear	ave_clustering_coeff_lcc	3.77E-05	3.82E-05	3.73E-05	all
0.1248096793	0.1172497175	0.1327026047	individual	linear	ave_clustering_coeff_lcc	1.40E-05	1.94E-05	8.51E-06	archaea
0.09020213201	0.08942968967	0.09098594656	individual	linear	ave_clustering_coeff_lcc	3.79E-05	3.83E-05	3.75E-05	bacteria
0.1020962312	0.08924806815	0.1181610154	individual	linear	ave_clustering_coeff_lcc	4.22E-05	4.76E-05	3.60E-05	eukarya
-7.190443076	-7.244080599	-7.135813277	ranRxn_individual	powerlaw	ave_clustering_coeff_lcc	0.6402690658	0.646515648	0.6337497909	ranRxn_individual
-3.815630862	-4.030552979	-3.595648125	syn_ecosystem	powerlaw	ave_clustering_coeff_lcc	0.2920747808	0.3178179809	0.2656702928	all
0.1465611133	0.1458008761	0.1473259738	syn_individual	linear	ave_clustering_coeff_lcc	2.21E-05	2.23E-05	2.19E-05	all
4.488665789	4.461928843	4.516055365	ecosystem	linear	ave_degree_lcc	0.0005452424434	0.0005524169757	0.0005378723417	JGI
4.169956135	4.160440884	4.179592243	individual	linear	ave_degree_lcc	0.0006795871293	0.0006844628218	0.0006747357713	all
4.316196618	4.241304113	4.393987306	individual	linear	ave_degree_lcc	0.000612137081	0.0006673028847	0.0005531632113	archaea
4.158957706	4.148982852	4.168840882	individual	linear	ave_degree_lcc	0.0006845344008	0.0006894623393	0.0006796419352	bacteria
4.165699789	3.717547379	4.608899517	individual	linear	ave_degree_lcc	0.0007314197301	0.0009147256222	0.0005574909737	eukarya
3.647509899	3.641913636	3.653123557	ranRxn_individual	linear	ave_degree_lcc	0.0003522613713	0.0003536911561	0.0003508403161	ranRxn_individual
0.5490346513	0.3618753781	0.7474150822	syn_ecosystem	powerlaw	ave_degree_lcc	0.1638509013	0.1856358054	0.1406779587	all
5.037734181	5.028919844	5.046321727	syn_individual	linear	ave_degree_lcc	0.0003709253877	0.0003730694676	0.0003688734281	all
1.81524011	1.794360853	1.836016412	ecosystem	powerlaw	ave_shortest_path_length_lcc	-0.08422048919	-0.0816306453	-0.08676167857	JGI
2.072169359	2.060409332	2.083916254	individual	powerlaw	ave_shortest_path_length_lcc	-0.1179516357	-0.1163972991	-0.1194772248	all
1.63824713	1.555132136	1.718391144	individual	powerlaw	ave_shortest_path_length_lcc	-0.06163791821	-0.05014150789	-0.07311933084	archaea
2.131804212	2.120817592	2.142607806	individual	powerlaw	ave_shortest_path_length_lcc	-0.1257996939	-0.1243471542	-0.1272338367	bacteria
1.814103636	1.306257742	2.115769798	individual	powerlaw	ave_shortest_path_length_lcc	-0.08534171635	-0.02059579196	-0.1241236057	eukarya
1.662925867	1.651851365	1.674053055	ranRxn_individual	powerlaw	ave_shortest_path_length_lcc	-0.05439981021	-0.05303317558	-0.05578635389	ranRxn_individual
1.089655589	0.9590196937	1.219302042	syn_ecosystem	powerlaw	ave_shortest_path_length_lcc	0.003003309311	0.01825911805	-0.0124586992	all
1.37899377	1.370904545	1.387424155	syn_individual	powerlaw	ave_shortest_path_length_lcc	-0.02959813939	-0.02861887811	-0.03062638057	all
-7.482753919	-7.557499362	-7.408219833	ecosystem	powerlaw	nbr_ec	1.838449545	1.847519874	1.829380121	JGI
-3.371922104	-3.438542156	-3.304884462	individual	powerlaw	nbr_ec	1.294009276	1.30315703	1.285255519	all
1.929150163	1.598316778	2.228198103	individual	powerlaw	nbr_ec	0.5440910994	0.5874250147	0.5023431713	archaea
-3.19173524	-3.252611453	-3.130685572	individual	powerlaw	nbr_ec	1.270178027	1.27824967	1.262236563	bacteria
-5.324882394	-6.168185187	-4.390111911	individual	powerlaw	nbr_ec	1.569133609	1.676210559	1.443734174	eukarya
-0.8211351999	-0.8379795527	-0.8046490267	ecosystem	powerlaw	nbr_edges_lcc	1.243813761	1.245819206	1.241770482	JGI

-0.648360844	-0.6597761376	-0.6371400607	individual	powerlaw	nbr_edges_lcc	1.219314446	1.220800229	1.217901476	all
-0.3872105249	-0.4901558906	-0.273748425	individual	powerlaw	nbr_edges_lcc	1.185072189	1.199196917	1.169709648	archaea
-0.6494426206	-0.6596918541	-0.6390714717	individual	powerlaw	nbr_edges_lcc	1.219410206	1.220728916	1.218030409	bacteria
-1.000663809	-1.435923792	-0.5805854499	individual	powerlaw	nbr_edges_lcc	1.2682855	1.323584689	1.213711244	eukarya
-1.104301078	-1.119578929	-1.088787641	ranRxn_individual	powerlaw	nbr_edges_lcc	1.245940062	1.247810134	1.244122806	ranRxn_individual
0.07964556041	-0.03273361671	0.2016258751	syn_ecosystem	powerlaw	nbr_edges_lcc	1.137618738	1.150499743	1.122595102	all
-0.5051692175	-0.5119621071	-0.4984291016	syn_individual	powerlaw	nbr_edges_lcc	1.203785287	1.204577665	1.202988505	all
-2.542243888	-2.564455158	-2.519415832	ecosystem	powerlaw	nbr_rxn	1.319772651	1.32251581	1.317001746	JGI
-1.86827931	-1.883225158	-1.853312026	individual	powerlaw	nbr_rxn	1.229494885	1.231434296	1.227495112	all
-1.252784933	-1.333655287	-1.172387776	individual	powerlaw	nbr_rxn	1.144161148	1.155104172	1.133007571	archaea
-1.853543987	-1.868385848	-1.838386665	individual	powerlaw	nbr_rxn	1.227518198	1.229483376	1.225619214	bacteria
-2.719131619	-3.052474626	-2.378718193	individual	powerlaw	nbr_rxn	1.342993444	1.38659353	1.298149597	eukarya
-3.174019772	-3.194913858	-3.153404429	ranRxn_individual	powerlaw	nbr_rxn	1.35908598	1.361531312	1.356672602	ranRxn_individual
-1.021832355	-1.181723618	-0.8408302368	syn_ecosystem	powerlaw	nbr_rxn	1.139751689	1.158156047	1.118757052	all
-2.003291852	-2.009127544	-1.997477283	syn_individual	powerlaw	nbr_rxn	1.253612529	1.254296954	1.252941754	all

Table S6: Distinguishability of synthetic ecosystem networks (randomly genome networks) scaling from that of real ecosystem.

Network Measure	P-value of permutation test
Number of Reactions	10^{-5}
Number of Edges (LCC)	10^{-5}
Average Degree (LCC)	10^{-5}
Average Shortest Path Length (LCC)	10^{-5}
Ave Clustering Coefficient (LCC)	10^{-5}
Average Node Betweenness	0.14
Average Edge Betweenness	0.08
Attribute Assortativity	0.256
Assortativity	0.210