Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/5/1/eaau0149/DC1

Supplementary Materials for

Universal scaling across biochemical networks on Earth

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

*Corresponding author. Email: sara.i.walker@asu.edu

Published 16 January 2019, *Sci. Adv.* **5**, eaau0149 (2019)
DOI: 10.1126/sciadv.aau0149

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

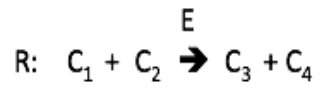
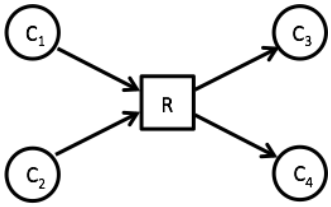
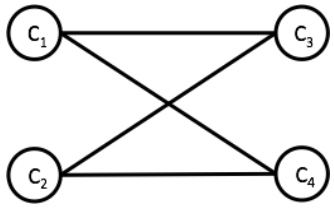
(available at advances.sciencemag.org/cgi/content/full/5/1/eaau0149/DC1)

Data file S2B (.csv format). Summary of measured network properties, by levels (parsed data only).
Data file S2C (.csv format). Summary of measured network properties, by levels (parsed data excluded).

Supplementary Materials

Section S1. Network representations of catalyzed 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. In principle, this biochemical reaction, R, can happen in opposite direction depending on the environment. Therefore, in bipartite network representation, the edges connecting chemical compounds and the reactions are considered as bidirected, which is equivalent to undirected for our analysis. (c) The unipartite network representation of the reaction, R, shows how the reaction information is embedded in the network. In the unipartite 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.

(a) Catalyzed Reaction	(b) Bipartite Network	(c) Unipartite Network
		

Section S2. Topological measures

To characterize the structure of biochemical networks, we utilized some of the most frequently used topological measures. The detailed descriptions about these topological measures can be found in[38]. Here, we briefly review these measures.. For computing each measure, we used the Python software package, NetworkX[73]. The topological measures implemented in this paper include average degree, average clustering coefficient, average shortest path length, assortativity (degree correlation coefficient), and node betweenness.

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. 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 component for each network generally contains 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.

Degree
The degree of a node *i*, k_i is the total number of connections between *i* and rest of the network. The average degree <*k*> in this paper is the average of k_i for all nodes in the LCC of a given network.

Clustering coefficient
The local clustering coefficient for a node *i*, C_i measures the local density of edges in a network by considering the number of connected pairs of neighbors of *i*. Hence, C_i is defined as,

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

where k_i is the degree of node *i* and L_i is the number of connections between neighbors of *i*. A large value of C_i indicates the highly interconnected neighborhood of *i*. The variable C_i is measured by using a Networkx method **clustering(..)**. We computed <*C*>, the average of C_i, over all nodes in the LCC of each network.

Shortest path length
The shortest path length, l_{ij} between a given pair of two nodes *i* and *j* is defined as the minimum number of edges connecting the two nodes in a given network. The variable l_{ij} is computed using the Networkx method **shortest_path_length(..)**. We calculated the average shortest path length, <*l*> by averaging l_{ij} for every pair of nodes in LCC of a given network.

Assortativity (degree correlation coefficient)
Assortativity measures the tendency of two nodes with similar properties to be connected in a given network. The assortativity coefficient proposed by Newman[74] is formulated as follows:

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b} \tag{2}$$

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. The variables σ_a and σ_b are the standard deviations of the distributions of a_x and b_y. When the considered attribute of nodes is their degree, the assortativity becomes the degree correlation coefficient, quantifying the correlation between the degrees of nodes on either side of an edge. Hence, for undirected networks in our study, $a_x = b_y$ and $\sigma_a \sigma_b = \sigma^2$. 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 an arbitrary network, -1<= *r* <= 1. To measure the assortativity *r* , we used a Networkx method **degree_assortativity_coefficient(..)**.

Betweenness
Betweenness centrality of a node, B_i is defined as [75],

$$B_i = \sum_{s,t \in V} \frac{\sigma(s,t|i)}{\sigma(s,t)} \tag{3}$$

where V is the set of all nodes in a network, and $\sigma(s,t)$ and $\sigma(s,t|i)$ denote the number of all shortest paths from s to t , and the number of the shortest paths through a given node i, respectively. Replacing $\sigma(s,t|i)$ with $\sigma(s,t|e)$ for an edge e, one can also formulate the edge betweenness. The variable B_i measures degree of importance of i for the interactions between subsets of a given network. To compute B_i, Networkx methods **betweenness_centrality(..)** is implemented and $<B>$ is average of B_i over every node in LCC of a given network.

Supplementary Figures

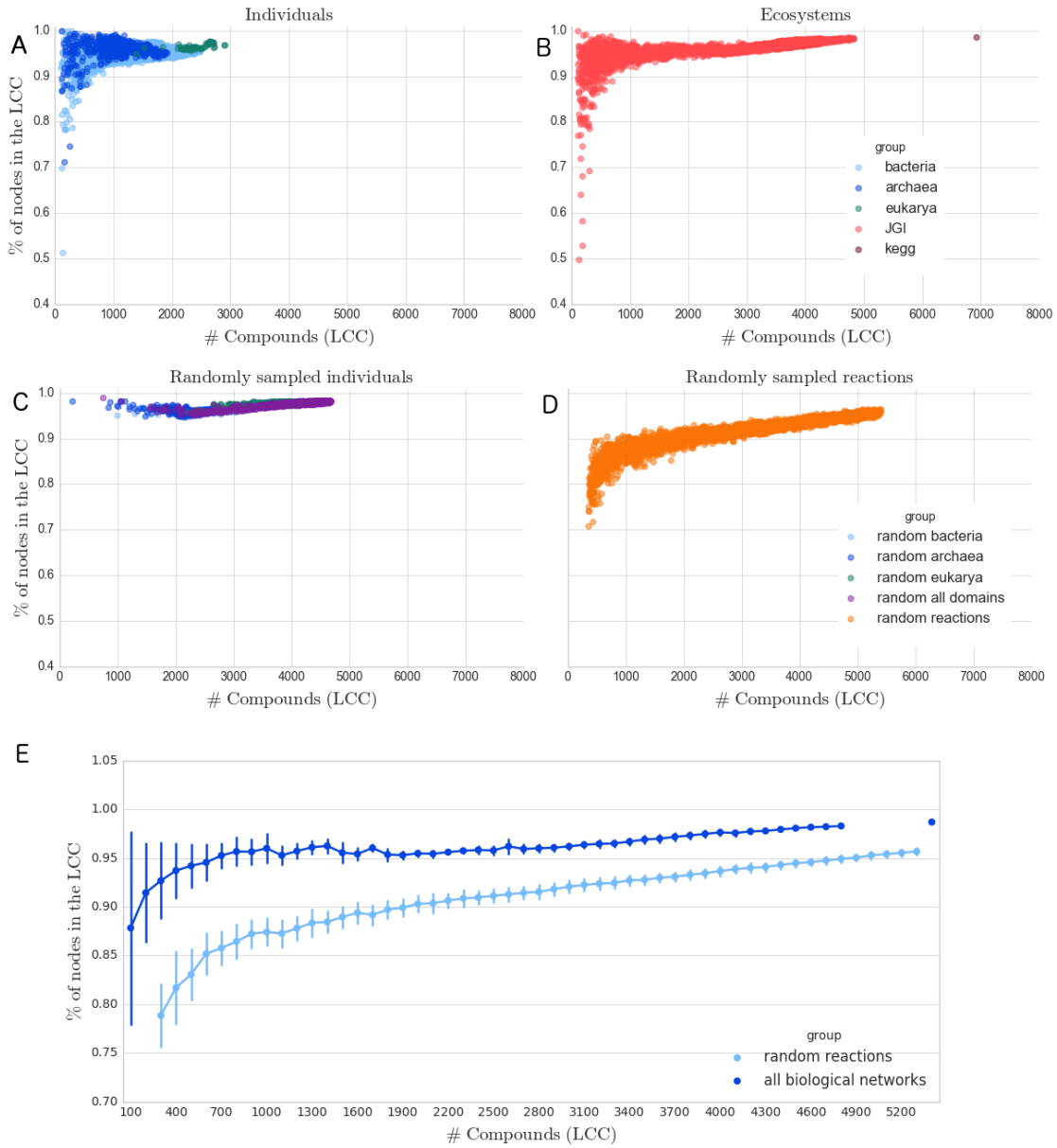


Fig. S1. Percentage of nodes in the LCC of a network versus the size of its LCC. Shown is the percentage of nodes in the LCC, as a function of the size of the network's largest connected component: **(A)** for all biological individuals (archaea, bacteria, eukarya). **(B)** for all biological ecosystems (from JGI, KEGG). **(C)** for randomly sampled individuals (archaea, bacteria, eukarya, and random individuals drawn from all domains). **(D)** for randomly sampled reactions. **(E)** Pointplot of biological networks (individuals and ecosystems) and random reaction networks, binned in increments of 100 compound nodes. Bars show one standard deviation of networks within a bin.

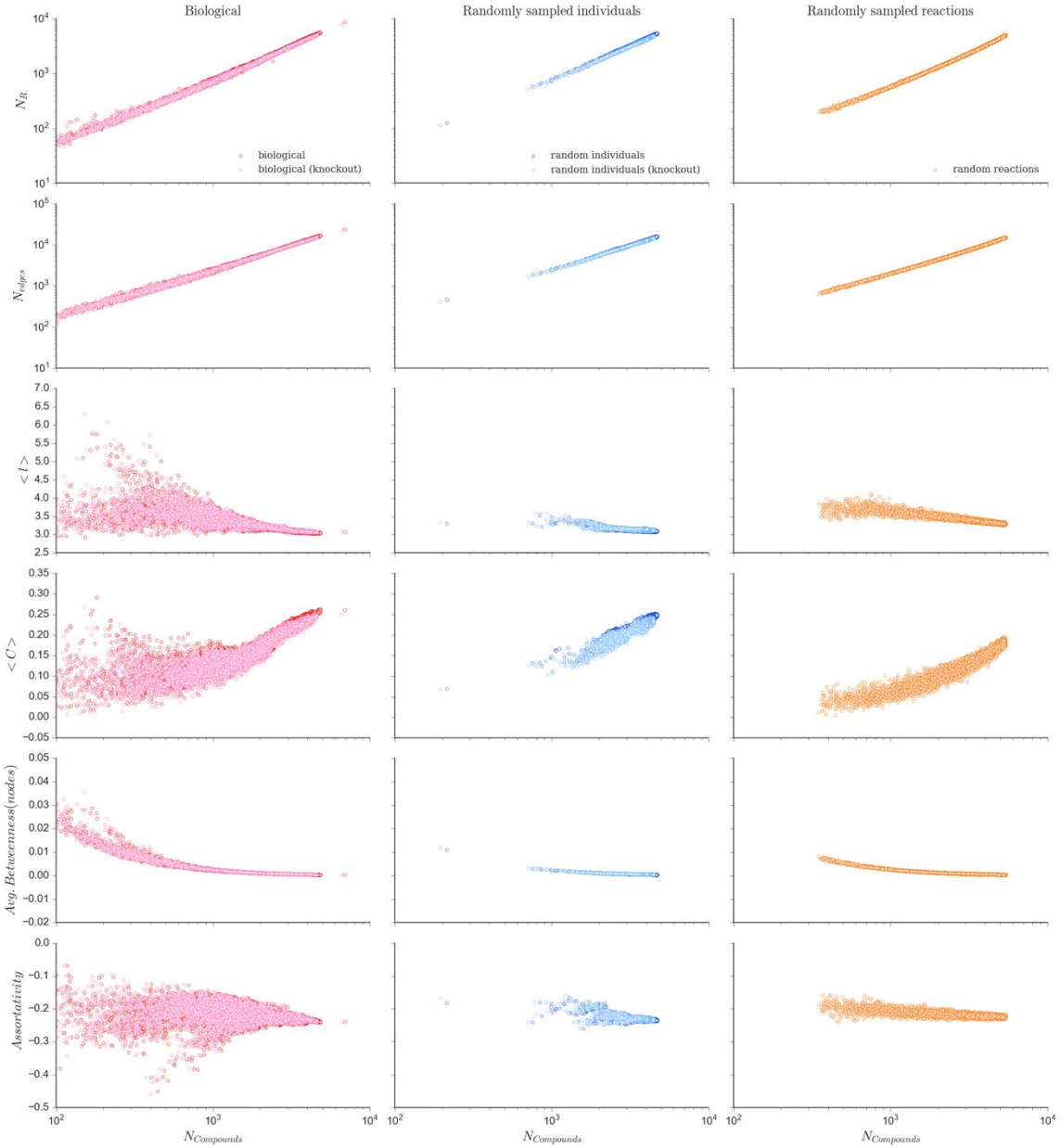


Fig. S2. Reaction knockout for unipartite networks. Diversity and topological measures shown for biological networks (left column), randomly sampled individual networks (center column), and randomly sampled reaction networks (right column). Original networks (bold colors) are compared to networks in the same category with 10% of their reactions randomly removed (pale colors). Random reaction networks are shown for comparison, but do not have knocked-out reactions (and cannot, by nature of their construction). Network measure scaling trends are not impacted by the removal of 10% of reactions, indicating our results are robust to missing data. Rows from rom top to bottom show: number of reactions (N_R), number of edges (N_{Edges}), avg. shortest path length (<l>), avg. clustering coefficient (<C>), avg. betweenness of nodes (<B>), assortativity (r).

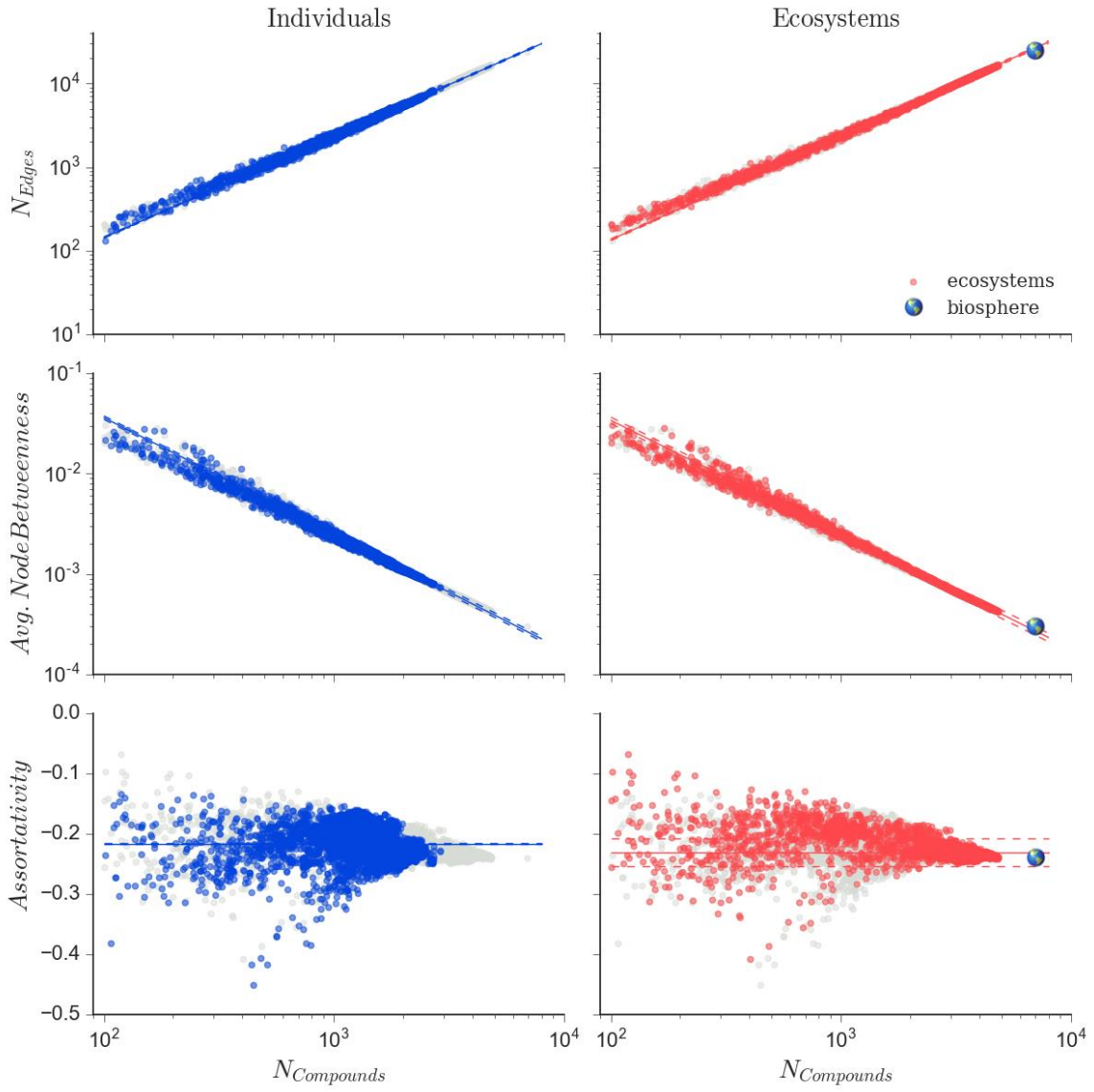


Fig. S3. Additional network measures for individuals and ecosystems show universal scaling across levels. Scaling behavior for additional topological measures for unipartite networks, to what is shown in main text Fig. 3. From top to bottom, number of edges (N_{Edges}), average node betweenness (<B>), assortativity (r).

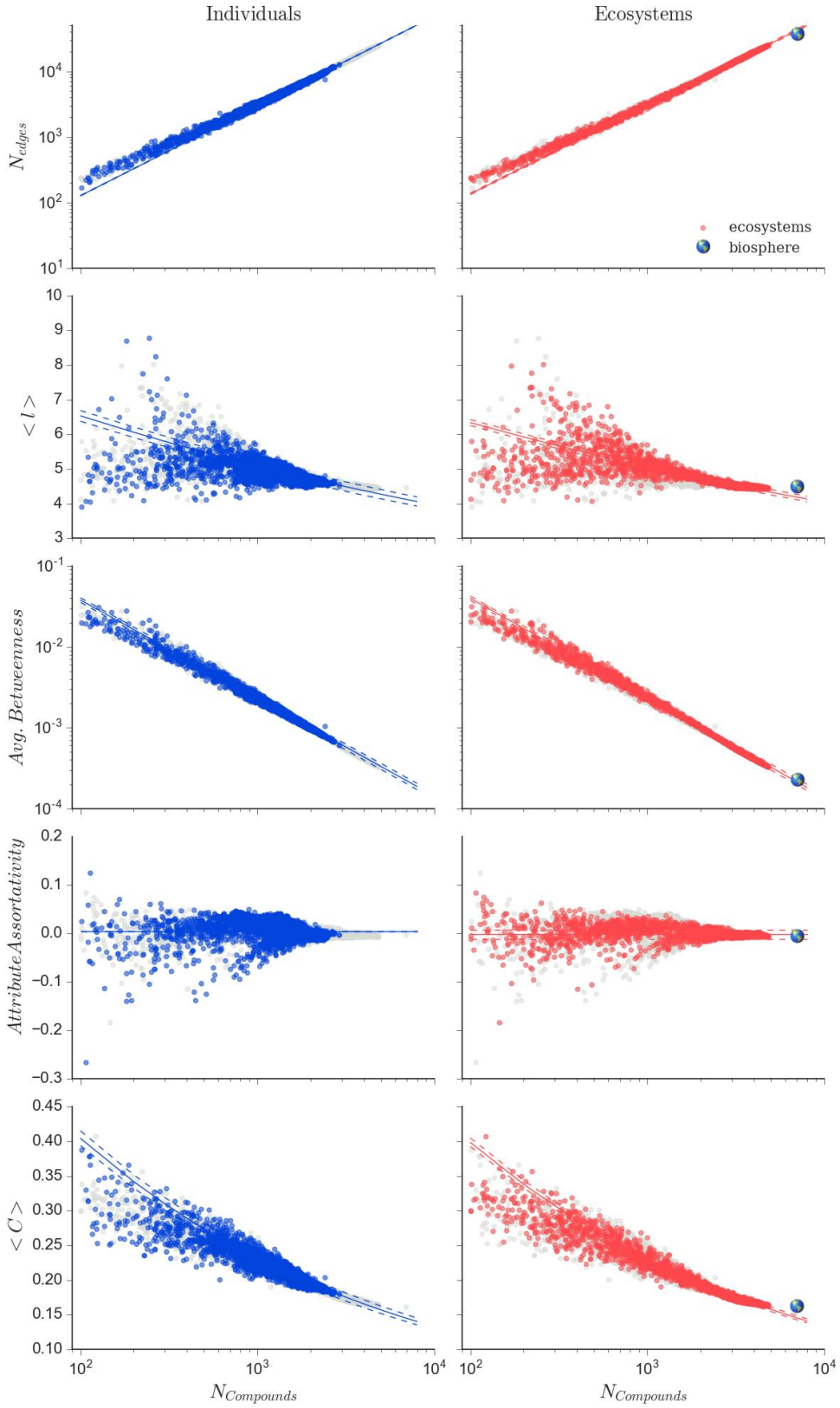


Fig. S4. Scaling of bipartite network structure for individuals and ecosystems. Shown are topological measures for bipartite representations of biochemical networks for individuals and ecosystems.. Our results demonstrate universal scaling behavior across levels is consistent across both unipartite and bipartite representations. Rows from top to bottom show number of edges (N_{Edges}), average shortest path length (<*l*>), average node betweenness (<B>), assortativity (r), and average clustering coefficient (<*C*>).

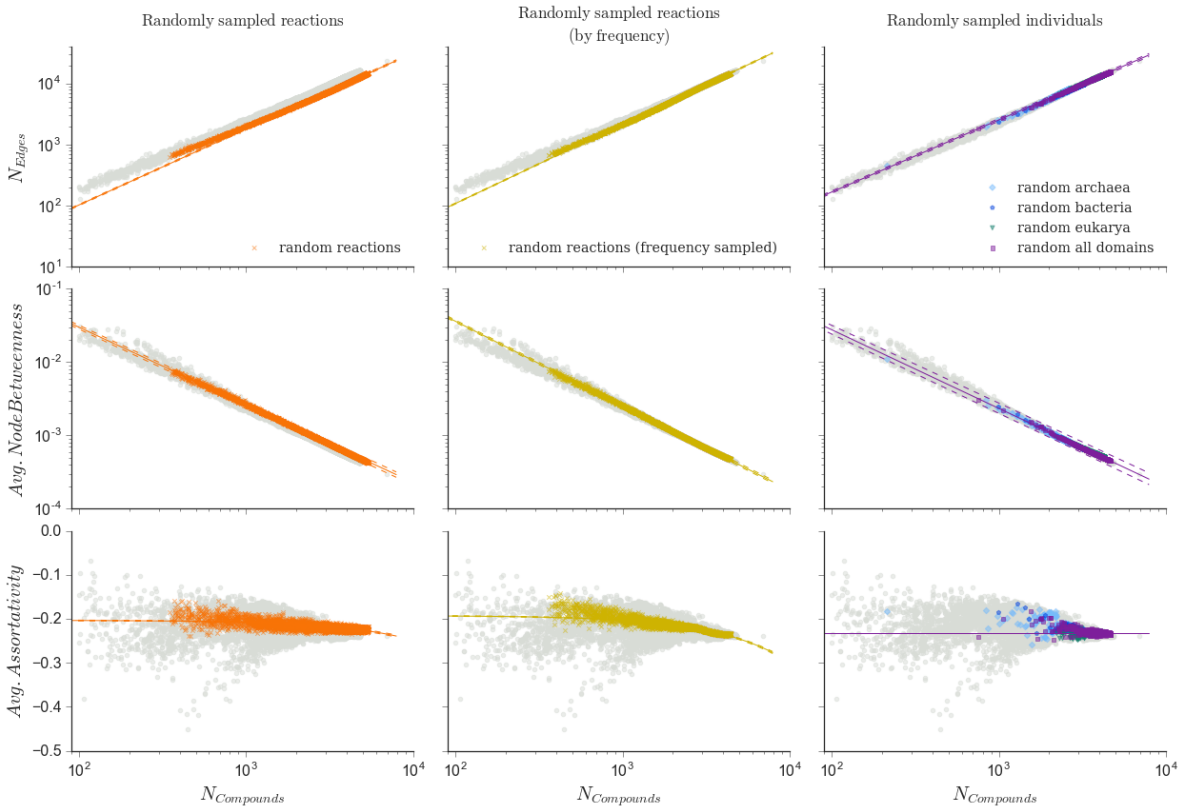


Fig. S5. Additional network measures for randomly sampled individuals and randomly sampled reactions. Scaling behavior for additional topological measures to those shown in main text Fig. 4. From top to bottom, number of edges (N_{Edges}), average node betweenness (<B>), and assortativity (r).

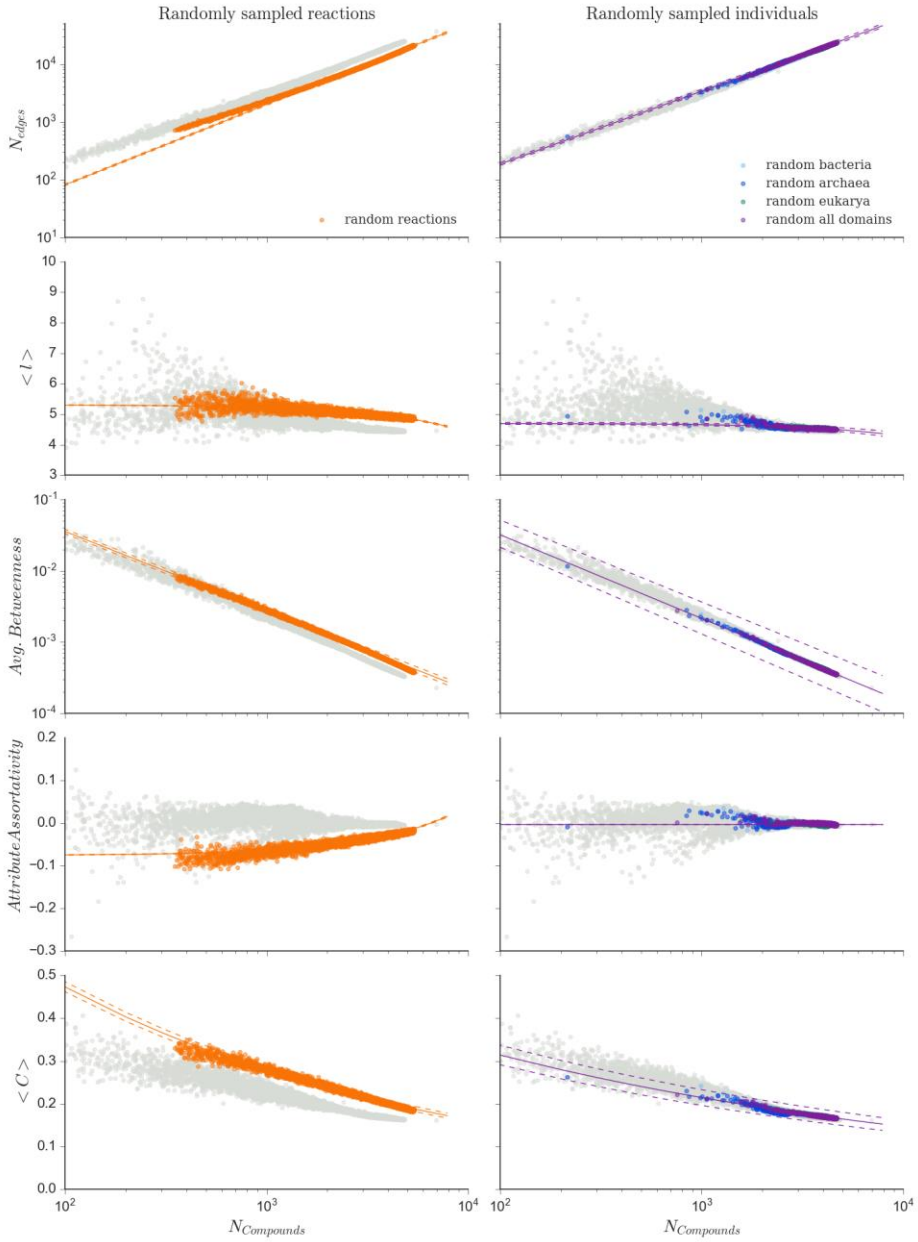


Fig. S6. Scaling of bipartite network structure for randomly sampled individuals and randomly sampled reactions. Shown are topological measures for bipartite representations of the random reaction networks and random genome networks. Our results show consistent scaling behavior in comparing the different data sets for both unipartite and bipartite representations. Rows from top to bottom, number of edges (N_{Edges}), average shortest path length (<*l*>), average node betweenness, assortativity (r), and average clustering coefficient (<*C*>).

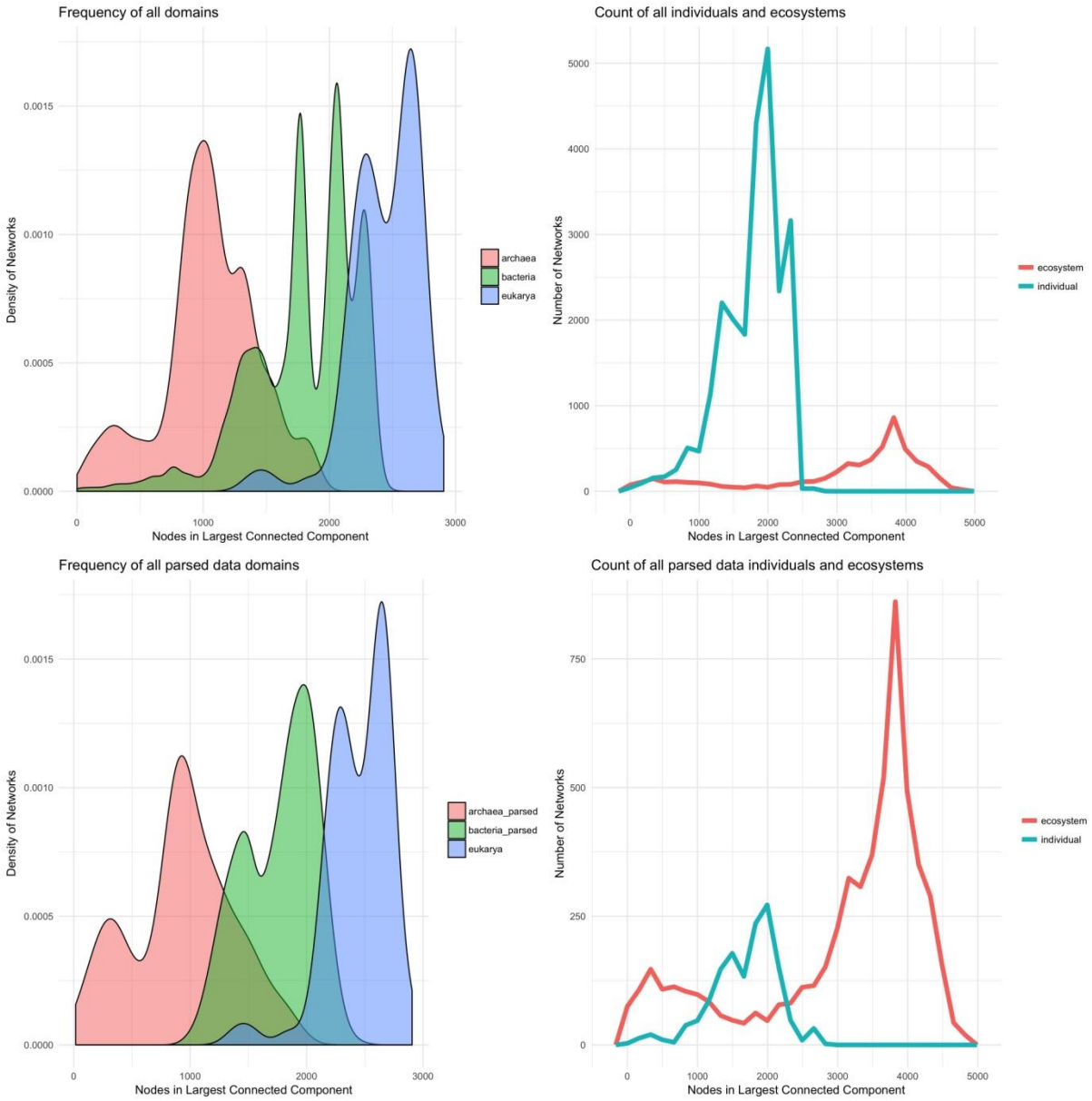


Fig. S7. Distributions of network sizes for each domain and across levels of organization. Top row: The relative distribution of network sizes (normalized to 1 over all networks of a given type) for networks in each domain (left), and the total number of networks in individuals and ecosystems (right). Bottom row: The relative distribution of network sizes for networks in each domain, for parsed datasets (left), and the total number of networks in individuals and ecosystems, for parsed datasets (right).

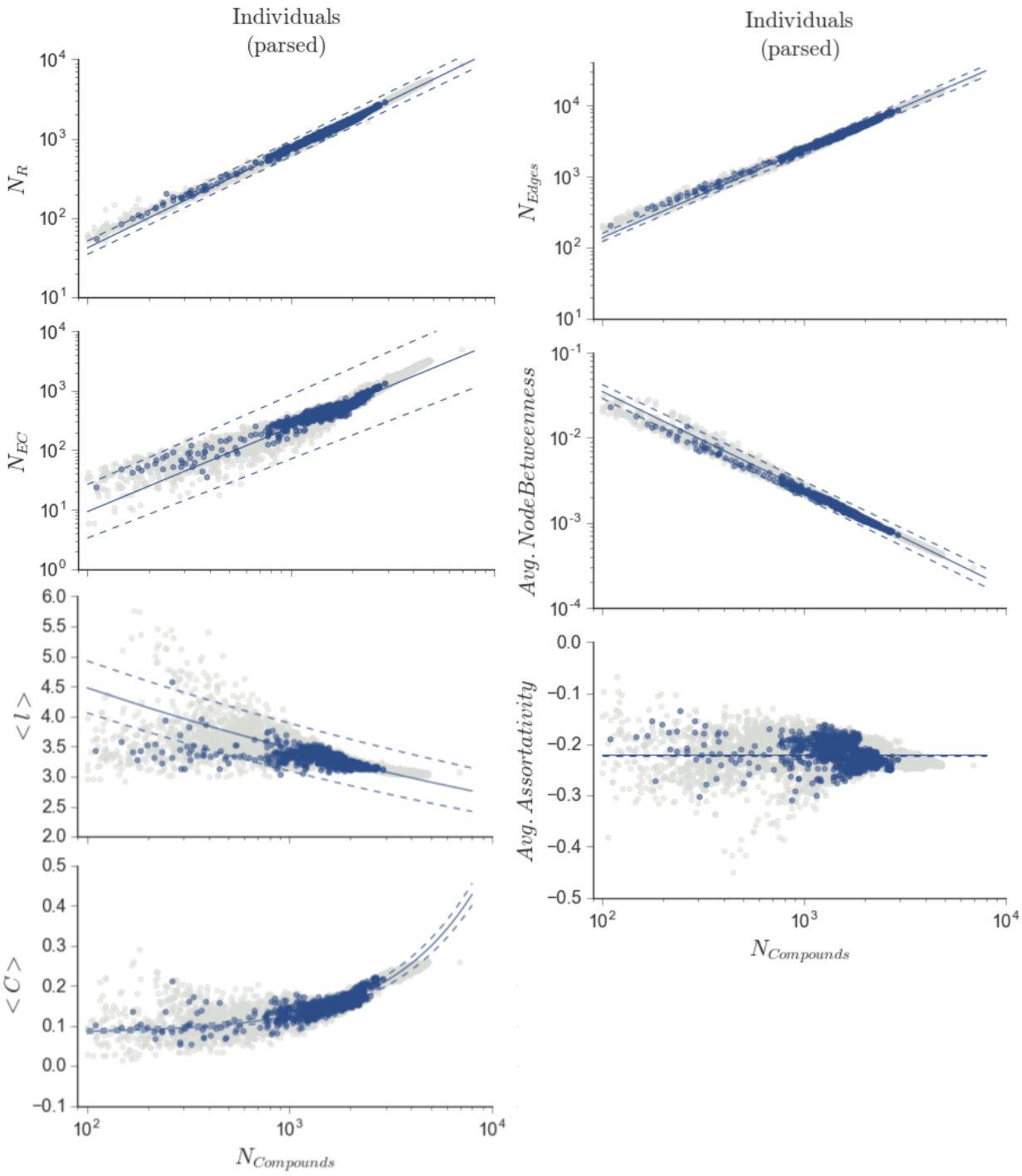


Fig. S8. Biochemical diversity and network topology measures for parsed datasets. Shown are data for unipartite representations of the parsed networks we analyzed. Left column, from top to bottom: number of reactions (N_R), number of ECs (N_{EC}), average shortest path length (<*l*>), and average clustering coefficient (<*C*>). Right column, from top to bottom: number of edges (N_{Edges}), average node betweenness (<B>), and assortativity (r).

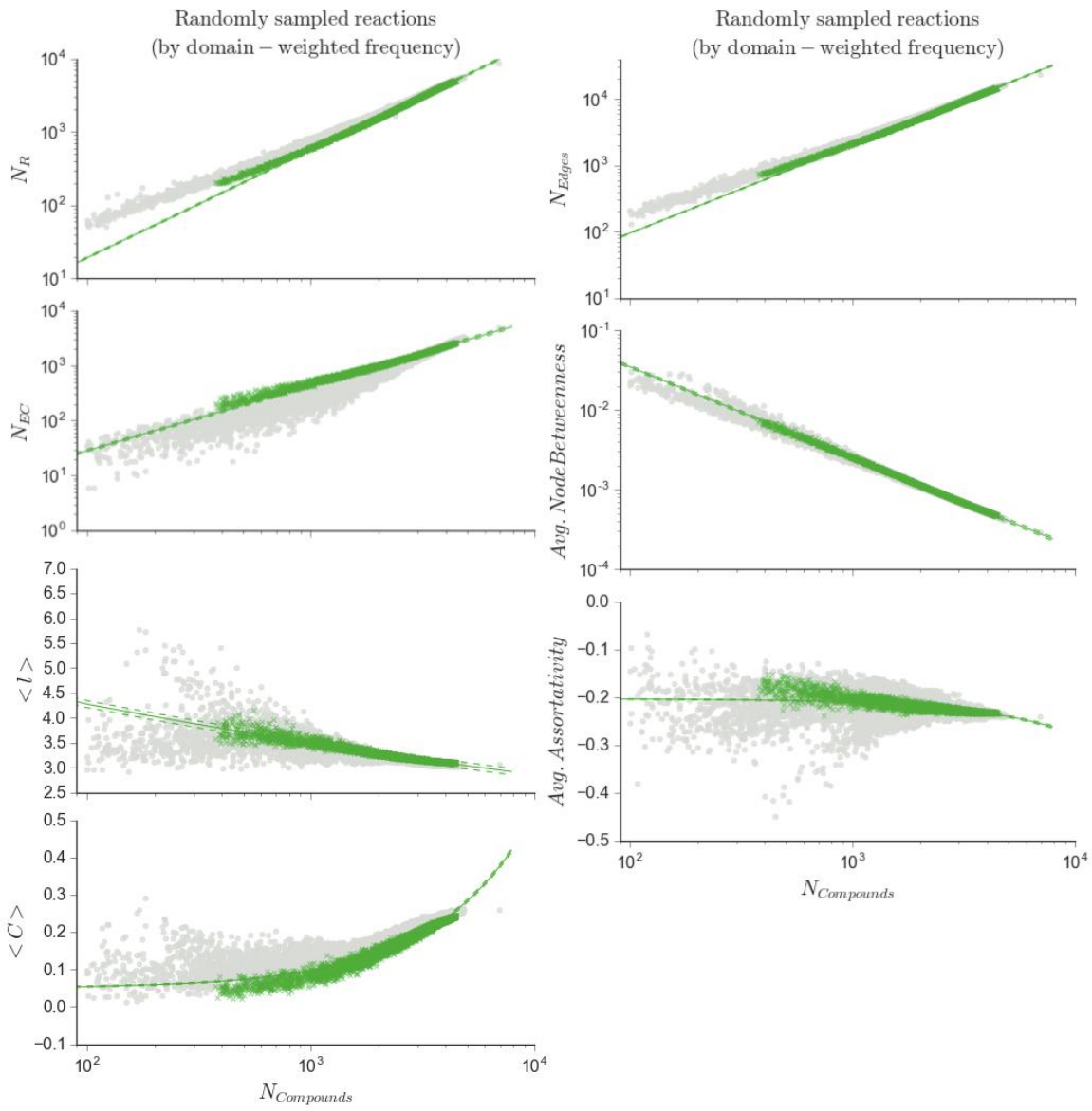


Fig. S9. Biochemical diversity and network topology measures for domain-weighted frequency-sampled random reaction networks. Shown are data for unipartite representations of the domain-weighted frequency-sampled random reaction networks we analyzed. These were created by sampling reactions based on the frequency distribution observed within each domain, with reactions from each domain given an equal probability to be sampled. Left column, from top to bottom: number of reactions (N_R), number of ECs (N_{EC}), average shortest path length (<*l*>), and average clustering coefficient (<*C*>). Right column, from top to bottom: number of edges (N_{Edges}), average node betweenness (<B>), and assortativity (r).

Supplementary Tables

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

	group	>85%	>90%	>95%
Biological individuals and ecosystems	Archaea	99.17	97.75	86.39
	Bacteria	99.84	99.65	87.53
	Eukarya	100.00	100.00	98.70
	JGI	98.10	97.06	88.42
	KEGG	100.00	100.00	100.00
Random genome	Archaea	100.00	100.00	99.75
	Bacteria	100.00	100.00	100.00
	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. Distinguishability of individuals and ecosystems, and ecosystems and random genome networks.

Property	Distinguishable Levels of Organization (p- value)	Distinguishability of Ecosystems and Random Genome Networks (p-value)
Number of Reactions, N_R	Yes (10^{-6})	Yes (10^{-5})
Number of Enzyme classes, N_{EC}	Yes (10^{-6})	NA
Average Betweenness (nodes), <B>	No (0.272)	No (0.14)
Average Betweenness (edges), $<B_{Edges}>$	No (0.185)	No (0.08)
Number of Edges (LCC), N_{Edges}	Yes (10^{-6})	Yes (10^{-5})
Mean Degree (LCC), <k>	Yes (10^{-5})	Yes (10^{-5})
Mean Clustering Coefficient (LCC), <C>	Yes (0.00853)	Yes (10^{-5})
Average Shortest Path Length (LCC), <l>	No (0.26893)	Yes (10^{-5})
Assortativity (LCC), r	No (0.0761)	No (0.210)
Assortativity for bipartite graphs (LCC), $r_{bipartite}$	No (0.0563)	No (0.256)

Supplementary Data Files

Data file S1. Scaling parameters for topological measures with 95% confidence intervals. The file entitled supplementary_data_s1-scaling_data.csv contains information for the scaling laws described in the main text. These data describe how various network and enzymatic properties scale with network size (the number of nodes in the largest connected component). This file has 11 columns (plus an index column) which identify the parameters of the fits. Each row is a different fit and each column contains information about the fit. The column entitled '**y.var**' indicates which network/enzymatic measure is being compared to network size. The column entitled '**projection**' indicates whether the network measure was applied to the unipartite or bipartite graph representation. The column '**level**' indicates the biological level of organization, value of '*individual*' corresponds to a network constructed from genomic data, '*ecosystem*' indicates a network constructed from metagenomic data, '*ranRxn_individual*' indicates networks of random biochemical reactions, '*syn_individual_all*' indicates networks constructed from random combinations of individual networks, '*parsed'* indicates networks constructed from parsed datasets (except for eukarya), *'bio_rand_uni'* indicates networks of random biochemical reactions weighted by their occurrence across all individual datasets, *'bio_rand_domain'* indicates networks of random biochemical reactions weighted by their occurrence within

each domain, with each domain's reactions given an equal probability to be included . The column labeled '**group**' indicates which part of the data set was used, this column only matters for the '*individual*' level columns. A group value of '*bacteria*' indicates scaling values for bacterial networks, similarly for the other two domains. The column entitled '**scaling**' indicates how the measure scales with size, with '*powerlaw*' meaning that measure scales following a power law, while '*linear*' means the measure scales linearly. A value of '*mean*' in the scaling column is used to show the measure does not scale with size. The remaining 6 columns contain numerical values for the scaling fits and their 95% confidence intervals. The mathematical meaning of these values depends on the scaling behavior of that measure (i.e. the corresponding value in the 'scaling' column). The value of '**alpha**' is always related to how the measure changes with size, while '**beta**' is always related to the intercept. If the scaling behavior is linear, then the measure scales according to y.var ~ alpha*(size) + beta, such that alpha is the slope of the line and beta is the intercept. If the scaling behavior is a power law, then the measure scales according to y.var ~ exp(beta)*(size)^alpha, such that alpha is the scaling exponent and exp(beta) is the intercept. The 95% confidence intervals have the same interpretation with the '**alphaP**' column indicating the upper bound on alpha and the '**alphaM**' column indicating the lower bound on alpha, the same convention is used for '**betaP**' and '**betaM**'. Measures that do not scale with size have values of zero in the alpha column, and the mean value is given in the beta column, with 95% of the distribution falling between betaM and betaP.

Data file S2A. Summary of measured network properties, by domain. The file entitled "supplementary_data_s2a-network_summary-groups.csv" contains a statistical summary of network properties, grouped by domain. Each row in the first column contains the name of the partition of data being described in that row, with "JGI" indicating the metagenomic data. Each column in the first row identifies the property which is being summarized in the rows below. The properties are as follows: **'nbr_rxn'** is the number of reactions encoded by the genome/metagenome; **'nbr_nodes'** is the number of nodes in the network; **'nbr_edges'** is the number of edges in the network; **'nbr_connected_components'** is the number of connected components in the network; **'nbr_nodes_lcc'** is the number of nodes in the largest connected component (LCC) of the network; **'nbr_edges_lcc'** is the number of edges in the LCC of the network; **'ave_degree_lcc'** is the average node degree in the LCC of the network; **'ave_clustering_coeff_lcc'** is the average clustering coefficient in the LCC of the network; **'ave_shortest_path_length_lcc'** is the average shortest path length in the LCC of the network; **'ave_betweenness_nodes_lcc'** is the average node betweenness in the LCC of the network; **'ave_betweenness_edges_lcc'** is the average edge betweenness in the LCC of the network; **'assortativity_lcc'** is the average assortativity in the LCC of the network; **'attribute_assortativity_lcc'** is the average attribute assortativity of the LCC of the network; **'diameter_lcc'** is the diameter of the LCC of the network; **'nbr_ecs'** is the number of enzyme commission numbers in the network. The statistical property being measured over all networks in a group, for a particular measure, are listed in each cell. The statistical properties are the count (number of networks), mean, std (standard deviation), minimum, maximum, and the quartiles.

Data file S2B. Summary of measured network properties, by levels (parsed data only). The file entitled "supplementary_data_s2b-network_summary-levels_parsed.csv" contains a statistical summary of network values, grouped together for the parsed networks (parsed archaea, parsed bacteria, and all eukarya). The format of the csv is otherwise identical to S2A (see description above).

Data file S2C. Summary of measured network properties, by levels (parsed data excluded). The file entitled “supplementary_data_s2c-network_summary-levels_noparsed.csv” contains a statistical summary of network values, grouped together by level for all data, excluding the parsed networks. “Individual” includes archaea, bacteria, and eukarya, and “ecosystem” includes all JGI metagenomic networks. The format of the csv is otherwise identical to S2A (see description above).