Supplementary Materials for
Universal scaling across biochemical networks on Earth

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

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

Supplementary Figures

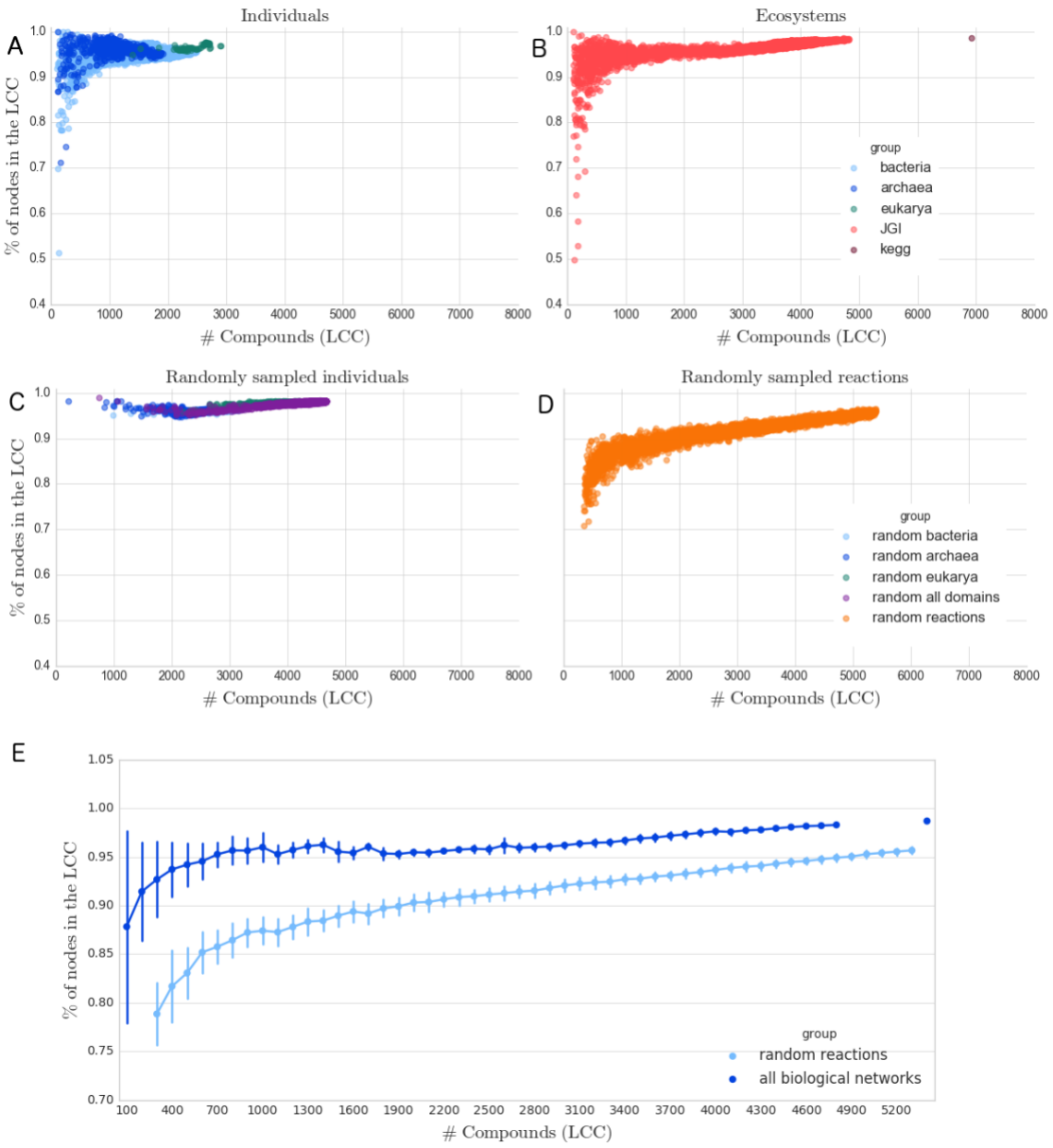

Fig. S1. Proportion of a network's nodes in the LCC versus its size.

The percent of nodes in each network's largest connected component (LCC), plotted as a function of the percentage of each network's nodes contained in its 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) vs. random reaction networks, binned in increments of 100 compound nodes. Bars show one standard deviation of networks within a bin.

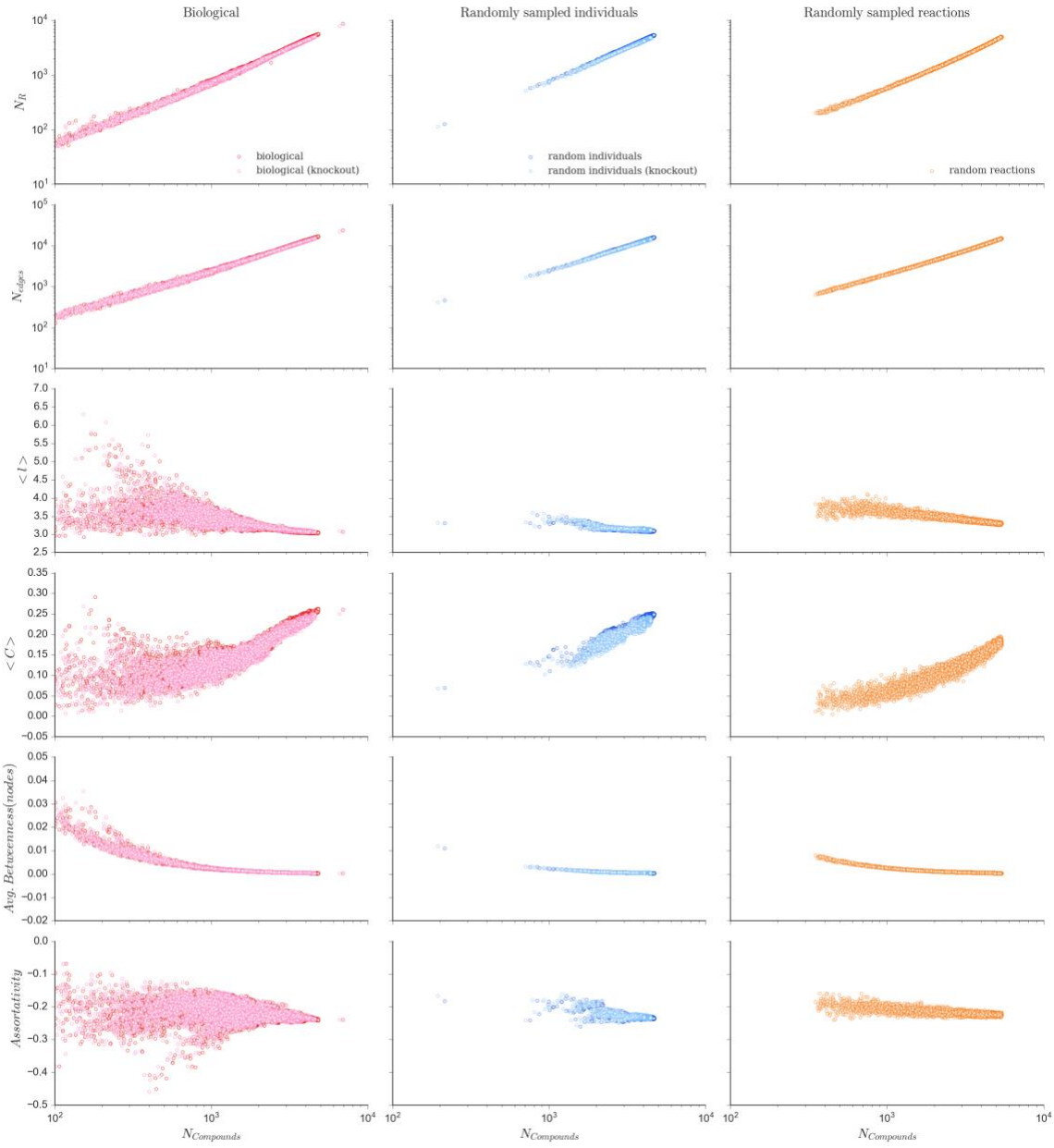

Fig. S2 Reaction knockout experiments for unipartite networks.

Network measure scaling trends are not impacted by the removal of 10% of reactions, indicating that our results are robust to missing data. From top to bottom: 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). Measures shown for biological networks (left column), randomly sampled individual networks (center column), and randomly sampled reaction networks (right column). Original networks are compared to networks in the same category that have had 10% of their reactions randomly removed. Random reaction networks are shown for comparison, but do not have knocked out reactions (and cannot, by nature of their construction).

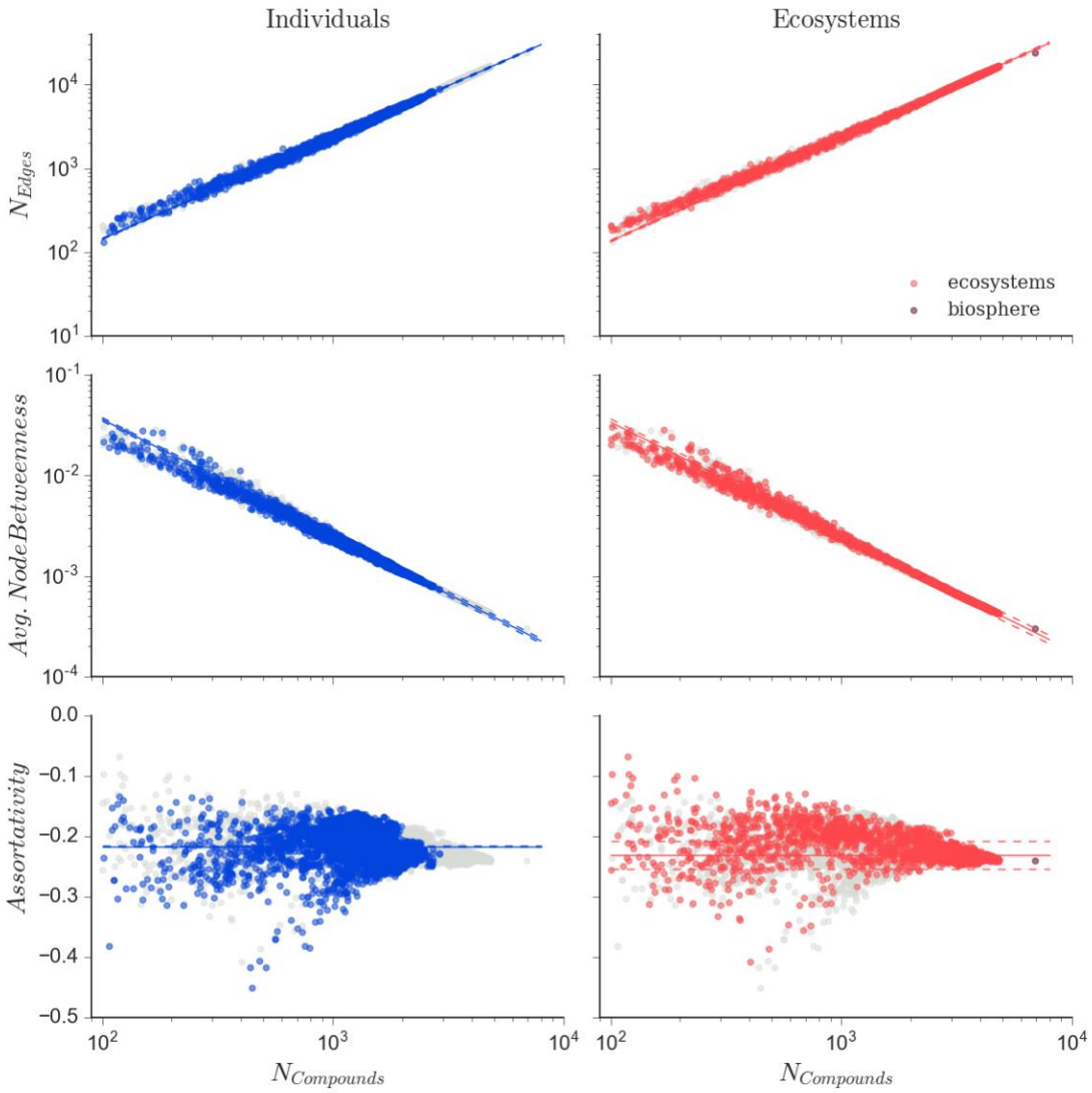

Fig. S3. Additional network measures for individuals and ecosystems (not shown in main text)

Scaling of biochemical networks for individuals (left column) and ecosystems (right column) for additional network topology measures to those shown in **Main text Fig. 3**. From top to bottom, number of edges (N_{Edges}), average node betweenness (<B>), assortativity (r).

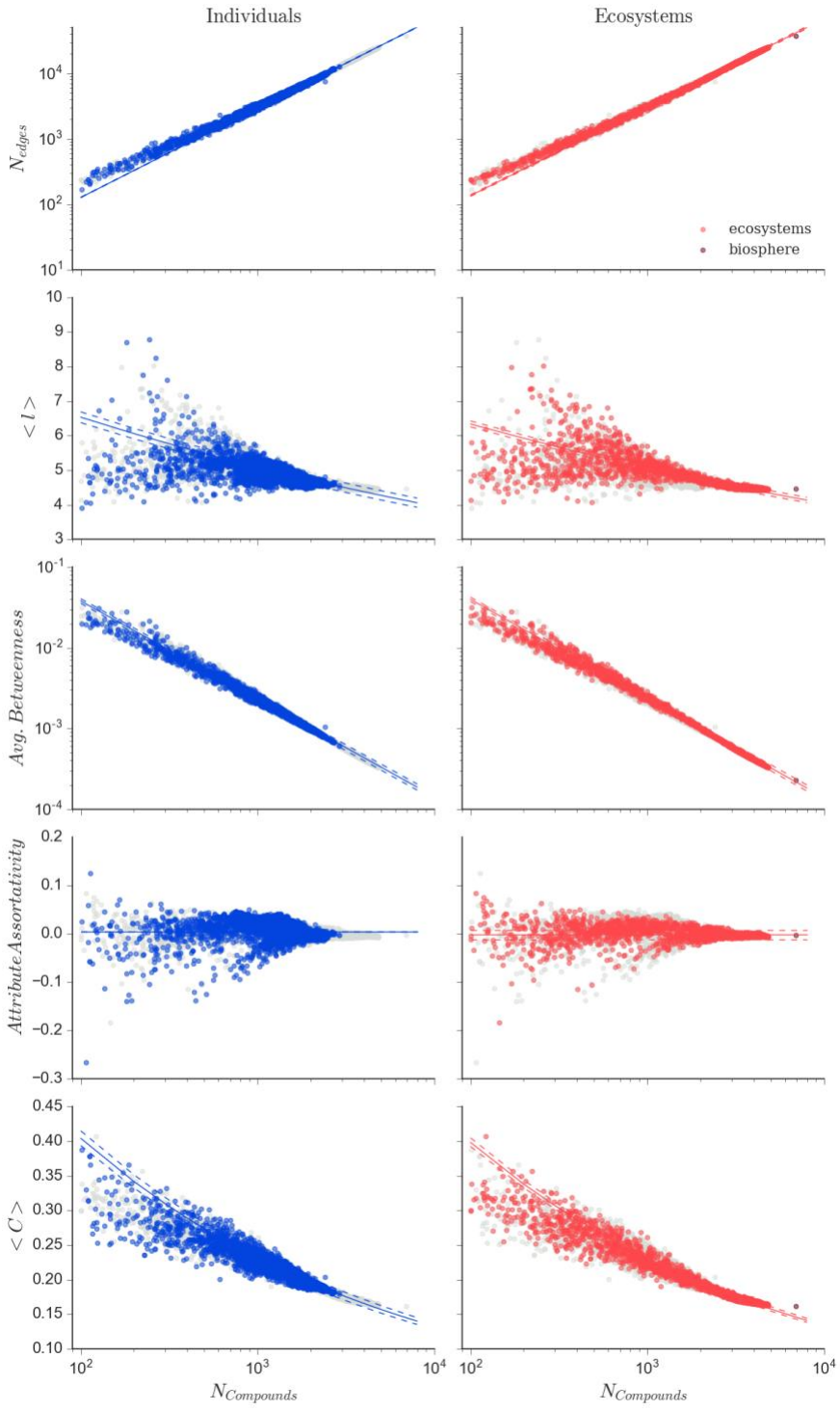

Fig. S4. Scaling behavior for bipartite network measures for individuals and ecosystems

Scaling of biochemical networks for individuals (left column) and ecosystems (right column) for bipartite representations. Results are consistent across both unipartite and bipartite representations in that both representations share universal scaling behavior across individuals and ecosystems. From top to bottom, number of edges (N_{Edges}),

average shortest path length (<*l*>), average node betweenness (<B>), assortativity (r), average clustering coefficient (<*C*>).

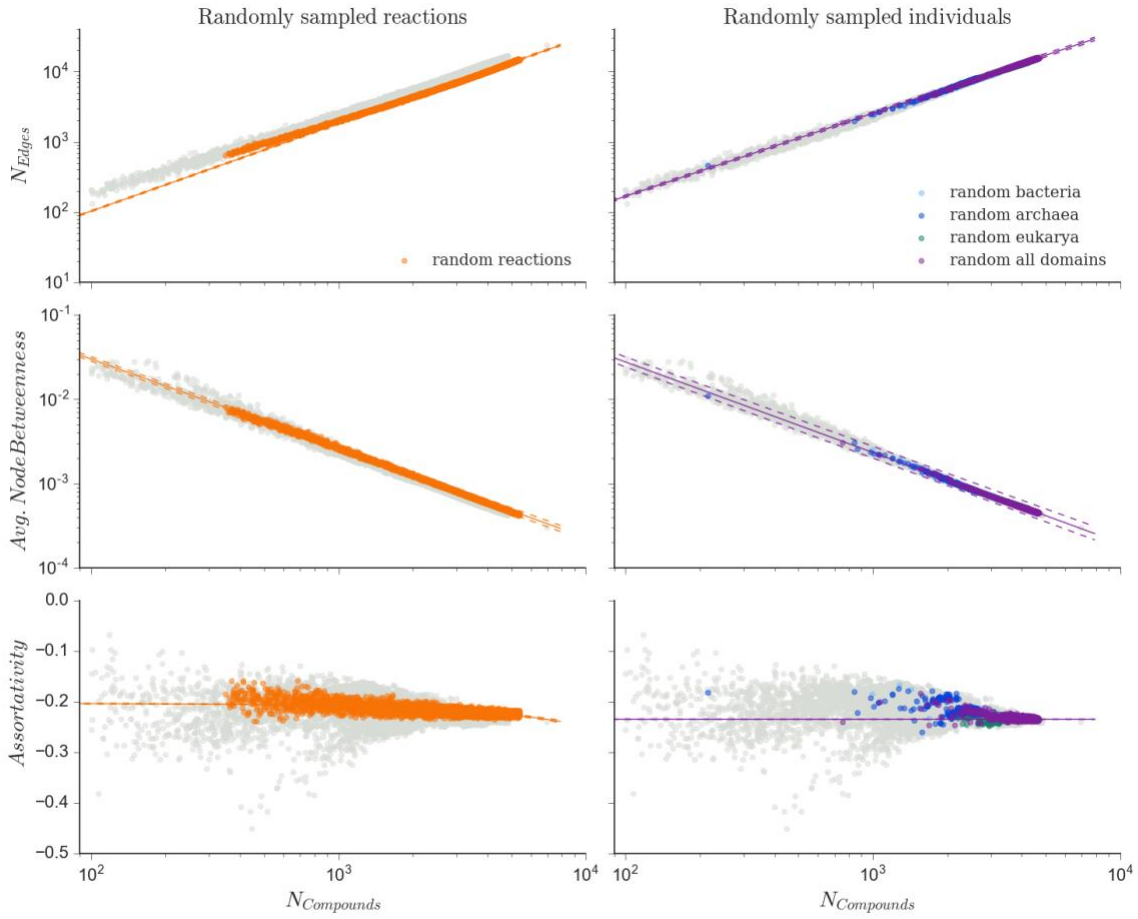

Fig. S5. Additional network measures for randomly sampled individuals and randomly sampled reactions (not shown in main text)

Scaling behavior for random reaction networks (left column) and random genome networks (right column) for additional network topology measures to those shown in **Main text Fig. 4**. From top to bottom, number of edges (N_{Edges}), average node betweenness (<B>), assortativity (r).

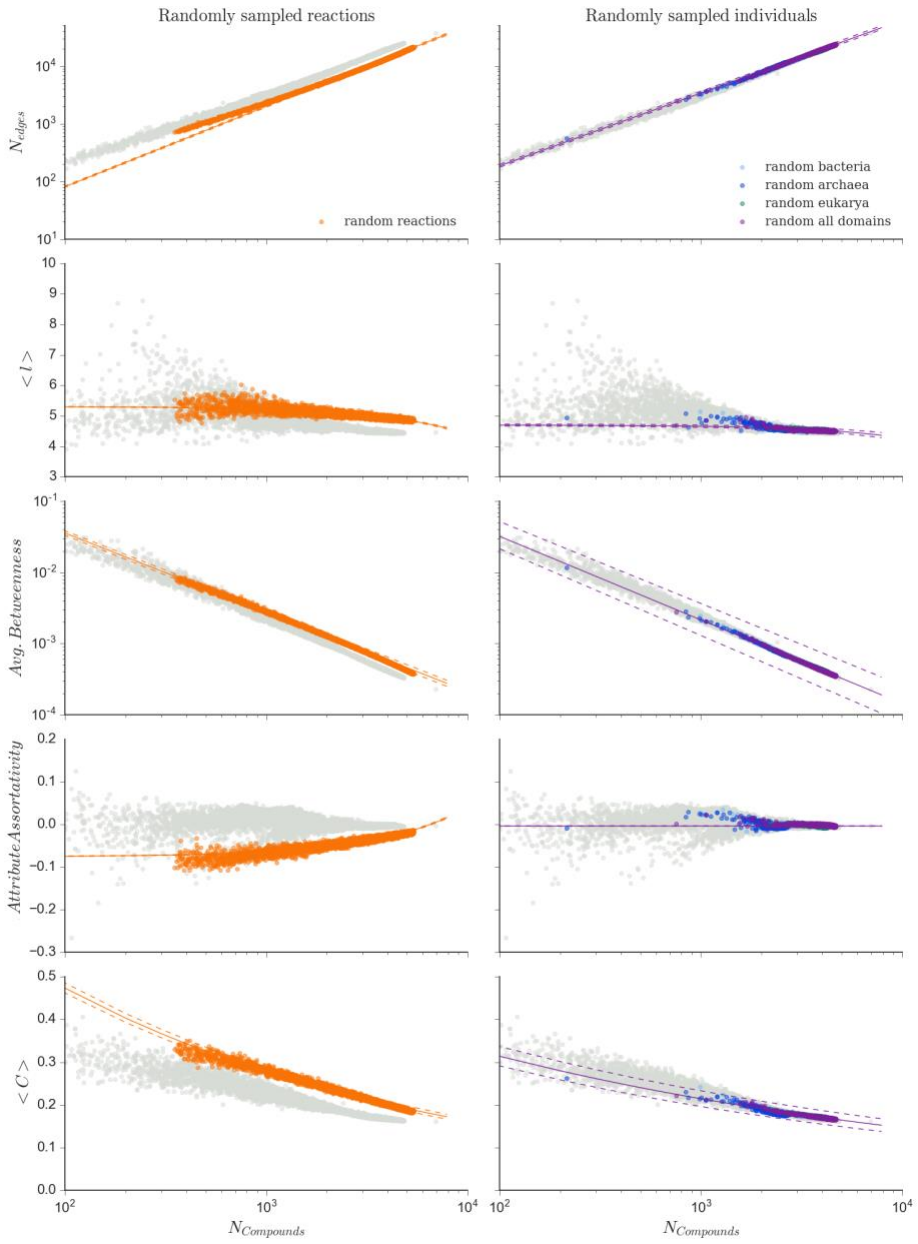

Fig. S6. Bipartite network measures for randomly sampled individuals and randomly sampled reactions

Scaling behavior for random reaction networks (left column) and random genome networks (right column) for bipartite representations. Results are consistent across both unipartite and bipartite representations in that both representations distinguish real biochemical networks from randomly generated ones. From top to bottom, number of edges (N_{Edges}), average shortest path length (<*l*>), average node betweenness, assortativity (r), average clustering coefficient (<*C*>).

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 ecosystem and of 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)

Data S1: *Scaling parameters for topological measures with 95% confidence intervals. See separate .csv for data. Description of Supplementary Data SI is below.*

The file entitled supplementary_data_s1-scaling_data.csv contains data for the scaling laws described in the main text. These data describe how various network and catalytic 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, where the value '*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, and '*syn_individual_all*' indicates networks constructed from random combinations of individual networks. 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 which 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.