Supplementary Material

Supplementary Discussion

Previous estimates of viral diversity

Our study re-analyzes the results of Carroll *et al.* [1], who used an existing approach that treats all viruses as host specific and all hosts as equivalent in diversity [2, 1, 3]. While our study estimates out to 100% host diversity first and then extrapolates missing viral diversity, Carroll *et al.* estimate missing viral diversity first. They begin by constructing rarefaction curves for two individual species as a function of sampling effort (viral species recorded versus individual animals tested); rarefaction curves are extrapolated from these data to give an estimate of the total number of virus species per host. These total diversity estimates are then used to calculate an estimate of the number of viruses in each family unique to each host species, and per-viral family per-host extrapolation is extended over the total diversity of mammals.

In the work of Anthony *et al.*, *Pteropus giganteus* was sampled to build rarefaction curves [2]; in the work of Carroll *et al.*, both *Pteropus giganteus* and *Macaca mulatta* were sampled [1], and estimates averaged across the two. Carroll *et al.*'s estimate of mammal diversity is broken down as

$$\begin{aligned} 1,531,745 \text{ viruses } &= 11.58 \text{ estimated viruses per family} \\ &\times 25 \text{ known viral families} \\ &\times 5,291 \text{ estimated mammal species} \end{aligned}$$

Notably, this methodology assumes no "forbidden links," i.e. connections are approximately equal rates between any host and virus relative to phylogeny, with no impossible or rare pairings.

Given the importance of influenza viruses from birds in human health, the authors similarly extend their estimate for birds using the same 11.58 viruses per family per host and one viral family (Orthomyxoviridae, the family of RNA viruses that include influenza), and add the estimate for birds (137,362.96) to their total to estimate 1,669,106 total viruses. In their dataset, 32.2% of viruses are zoonotic and 45.0% are human viruses, which they use as confidence bounds and multiply by 1.6 million to obtain an estimate of 631,218 to 826,647 zoonotic viruses.

Our parametric approach

Our study uses an approach first proposed by Strona and Fattorini (2014), which involves fitting a curve to data generated by iteratively subsampling a bipartite

host-parasite network. In their study, Strona and Fattorini propose that when a low proportion of hosts have been sampled, a power law is appropriate as an approximation of the curve [4]. Here, we briefly address the process of curve-fitting and the use of the power law.

In macroecology, disagreements about the use and misuse of power laws are encountered regularly [5, 6]. Perhaps the most notorious instance of this is the long history of scholarship examining the *species area relationship* (SAR), first described as a power law of form $S \sim A^{0.25}$ in 1921 [7]. In the century of research since, dozens of curves have since been posited as possible alternative formulations [8, 9, 10]. Controversy around these curves reflects curiosity and disagreement about the underlying mechanisms of scaling in natural systems, but also highlights the need for models that are predictive across scales. In the SAR literature, the latter is particularly important, as researchers often attempt to scale plot or landscape level data up to continental or global levels, spanning several orders of magnitude in both species and area [11, 12]. As macroecological theory has advanced, mechanistic formulations of the SAR, like those predicted by the neutral or maximum entropy theories of ecology, have become more common alternatives [13, 14, 15].

The scaling pattern we examine – a codependent richness relationship – lacks the benefit of a century's work on statistical underpinnings, but shares many problems with the SAR. Perhaps the most fundamental similarity is that both can be formulated in one of two ways. The *island* species area relationship uses data pulled from separate, independent geographic areas (such as islands in an archipelago), whereas the *nested* species area relationship considers the pattern of diversity accumulation in an expanding radius from one location (such as plots within a landscape). The scaling between host and affiliate richness can be thought of similarly. Previous work considering an island approach has found strong evidence for positive scaling patterns between host and parasite diversity [16], and a global study of host-helminth networks by country found a log-log linear scaling [17]. But so far we only know of one study (Strona and Fattorini's) that has taken the nested approach by subsampling bipartite networks.

Why a power law

Our decision to use a power law reflects a few major considerations. First, this approach is the only one with precedent using this data generation method [4]. Though this is no guarantee of adequacy, we suggest it makes it an appropriate starting point, especially as an incremental improvement on linear extrapolation. Second, for the six main networks in our paper (for which all results from Table S2-S6 and Figure S2-S4 are from the same 100 subsampled iterations), a power law outperforms a linear or logarithmic curve fit to the same data (Figure S2 and S3, Table S2), matching the findings of previous work. Moreover, in no case did the confidence intervals for the scaling exponents overlap 1, and all were within

the previously hypothesized 0.3–0.7 range. (Table S3)

Third, the power law is the simplest curve we considered that appears adequate to describe the broadest patterns in the data. The "true" pattern may be more complex, or not a power law at all; ultimately, this is not a biological question, but one of network statistics: under a certain set of conditions governing both degree distributions, what is the expected scaling between the number of edges on either side of a bipartite network randomly subsampled from a much larger one? As we have been unable to find any analytical solution to this problem in the network theoretic literature, we choose to default to the simplest (most parsimonious) model. The use of parsimony to constrain model selection is particularly common in instrumental applications of macroecology, as discussed at much greater length in a new paper by Coelho *et al.* [18]

A final consideration we have made revolves around the scale of extrapolation. Power laws are widely seen as "special" in the natural sciences due to their property of *self-similarity* (scale-invariance), which facilitates extrapolation over several orders of magnitude (as is common, for example, in species area relationship studies) [5]. When power laws are recognized in ecology, this is also sometimes taken as evidence of scale-invariance in the underlying biological *process*. Regardless of whether this is true, we argue it is largely irrelevant in our present case (and our instrumental use of a power law should not be taken as support of this assumption). With $\sim$ 500 mammals sampled of roughly 5,000 species globally, our models are only extrapolating over a single order of magnitude, and we would argue this largely circumvents the deeper question of scale invariance. Moreover, the tendency of the power law to overpredict at higher values as shown in Figure S2 (and the effect of predicting based on 10% of a network, shown in Figure 2) suggests that our estimates can be conservatively interpreted as an upper bound on possible diversity.

Alternatives to classical power laws

Plotting the residuals of the power law models show that the models perform most poorly at the lowest richness levels (Figure S4). These plots also illustrate that all six models overpredict at the lowest and highest values; the nonlinearity in the residuals might indicate that a more complicated functional form might better describe the results.

To explore these results further, we reviewed alternate scaling models that have been proposed in three papers from the species area relationship literature [8, 9, 10]. We limited candidate models to non-logarithmic models without indefinite polynomial expansions, and more importantly to those without an asymptote. The underlying logic of the second decision is that as long as a non-trivial proportion of species have host-specific affiliates, diversity should continue to accumulate as new hosts are added. The biological interpretation of this—that there should be

no level of host diversity where marginal returns of sampling new hosts approach zero—is an important one, as it highlights that our method proposes a conservative reduction from Carroll's previous viral diversity estimates. (The idea that viral diversity should asymptote at an intermediate level of host sampling would further reduce these estimates.)

We selected a total of five additional candidate models (Table S4). We used the same six simulated datasets, and used `nls` to fit additional models for five candidate models alongside the standard power curve. The AICs of the models are given in Table S5. Although the basic power law was not the best performing for any dataset, no model came out universally superior, though four of six were best fit by the quadratic expansion of the power law. We next examined the impact of these model differences on the DNA and RNA viral richness estimates by extrapolating all six candidate models to 5,291 mammal species. All expanded model forms produced dramatic reductions in the estimated viral richness (Table S6), especially because we did not constrain our models to monotonically positive forms; these differences that would only be further amplified after the sampling correction we apply in the main text. Given the wide range in estimates between models with no clear leader, and given that the classic power law method is both the simplest, easiest to interpret, and seems an adequate upper bound, we elected to use it in the main text analyses.

References

[1] Carroll, D., Daszak, P., Wolfe, N. D., Gao, G. F., Morel, C. M., Morzaria, S., Pablos-Méndez, A., Tomori, O., and Mazet, J. A. *Science* **359**(6378), 872–874 (2018).

[2] Anthony, S. J., Epstein, J. H., Murray, K. A., Navarrete-Macias, I., Zambrana-Torrelio, C. M., Solovyov, A., Ojeda-Flores, R., Arrigo, N. C., Islam, A., Khan, S. A., et al. *MBio* **4**(5), e00598–13 (2013).

[3] Olival, K. J., Hosseini, P. R., Zambrana-Torrelio, C., Ross, N., Bogich, T. L., and Daszak, P. *Nature* **546**(7660), 646–650 (2017).

[4] Strona, G. and Fattorini, S. *International Journal for Parasitology* **44**(5), 269–272 (2014).

[5] Brown, J. H., Gupta, V. K., Li, B.-L., Milne, B. T., Restrepo, C., and West, G. B. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **357**(1421), 619–626 (2002).

[6] Harte, J. and Kitzes, J. In *Saving a Million Species*, 73–86. Springer (2012).

[7] Arrhenius, O. *Journal of Ecology* **9**(1), 95–99 (1921).

[8] Dengler, J. *Journal of Biogeography* **36**(4), 728–744 (2009).

[9] Tjørve, E. *Journal of Biogeography* **30**(6), 827–835 (2003).

[10] Tjørve, E. *Journal of Biogeography* **36**(8), 1435–1445 (2009).

[11] Harte, J., Smith, A. B., and Storch, D. *Ecology Letters* **12**(8), 789–797 (2009).

[12] Harte, J., McCarthy, S., Taylor, K., Kinzig, A., and Fischer, M. L. *Oikos* **86**, 45–54 (1999).

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

[14] Hubbell, S. P. *The unified neutral theory of biodiversity and biogeography (Monographs in Population Biology 32)*. Princeton University Press, (2001).

[15] Rosindell, J. and Cornell, S. J. *Ecology Letters* **10**(7), 586–595 (2007).

[16] Kamiya, T., O'dwyer, K., Nakagawa, S., and Poulin, R. *Ecography* **37**(7), 689–697 (2014).

[17] Dallas, T. A., Aguirre, A. A., Budischak, S., Carlson, C., Ezenwa, V., Han, B., Huang, S., and Stephens, P. R. *Global Ecology and Biogeography* **27**(12), 1437–1447.

[18] Coelho, M. T. P., Diniz-Filho, J. A., and Rangel, T. F. *Ecography* (2018).

Supplementary Figures and Tables

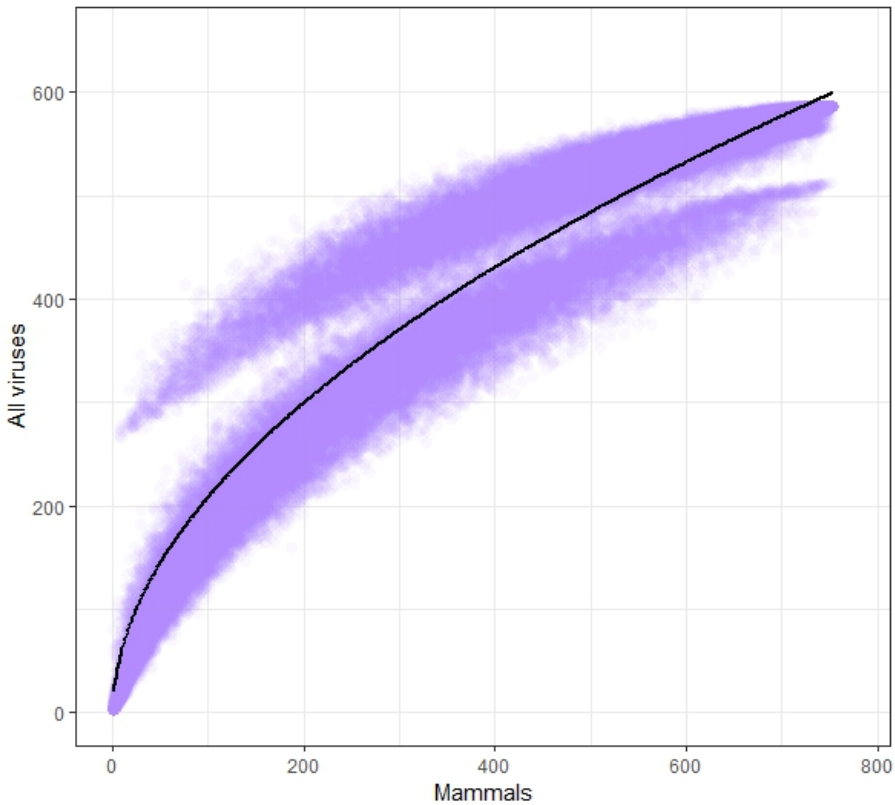


Figure S1: Bipartite rarefaction curves on the known viral network, if humans are included. Curves bifurcate due to the atypically high degree of *Homo sapiens*, leading to a poor fit in the overall model.

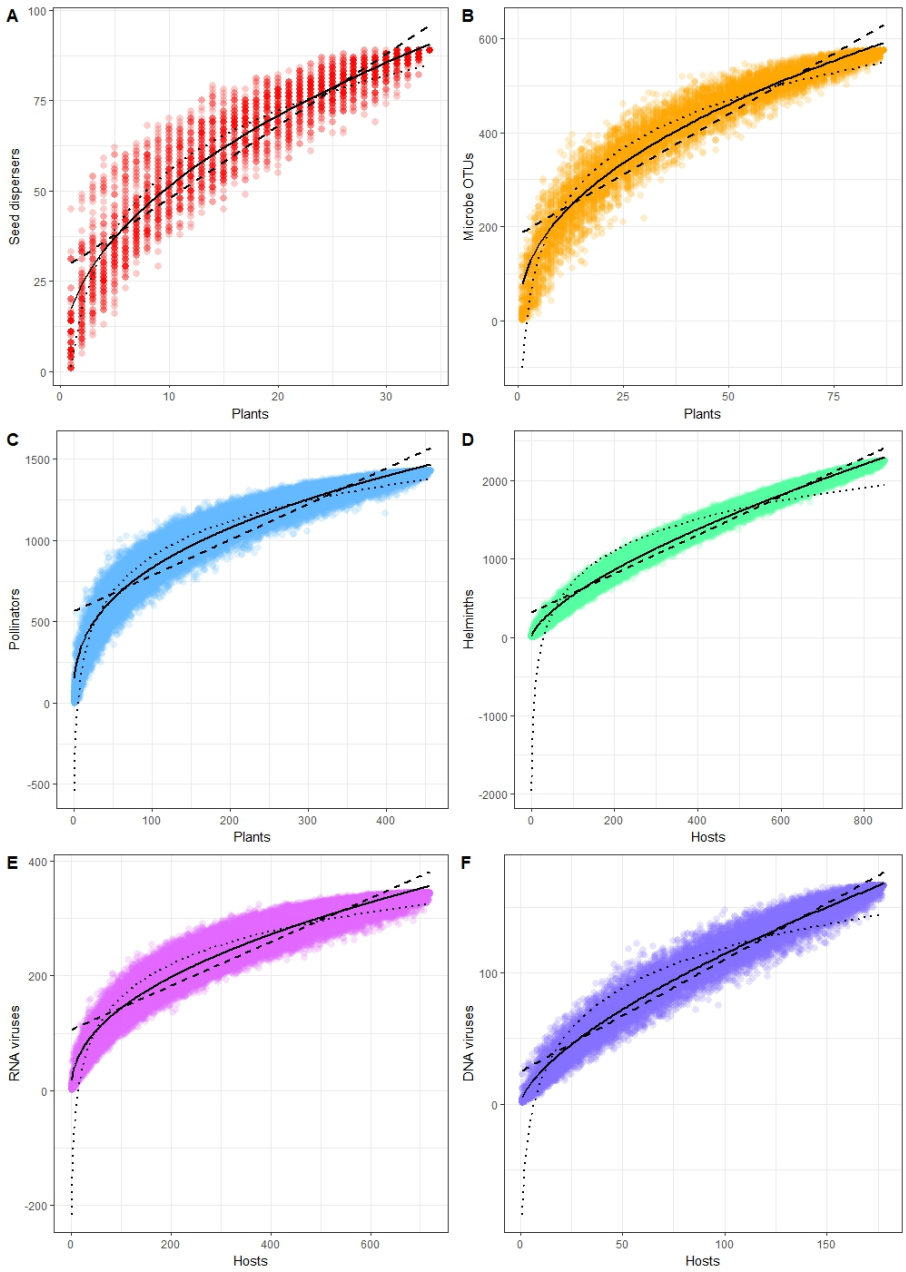


Figure S2: Three models fit to the six datasets: linear (long dash), power law (solid line), and log (dotted). These curves are all fit with 100 iterations generated in the subsampling procedure.

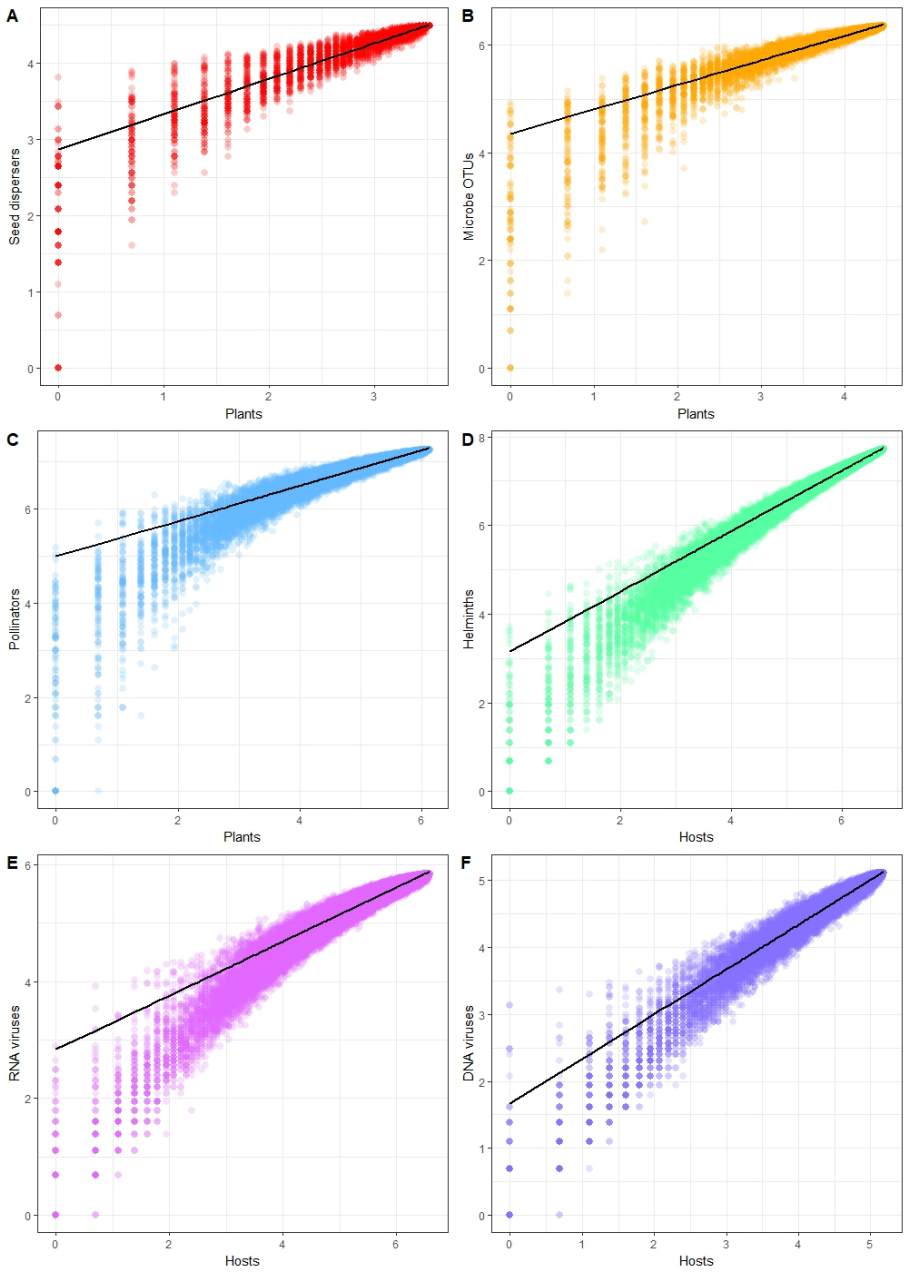


Figure S3: Log-log plots of the six main experimental networks we discuss in the main text. These curves are all fit with 100 iterations generated in the subsampling procedure.

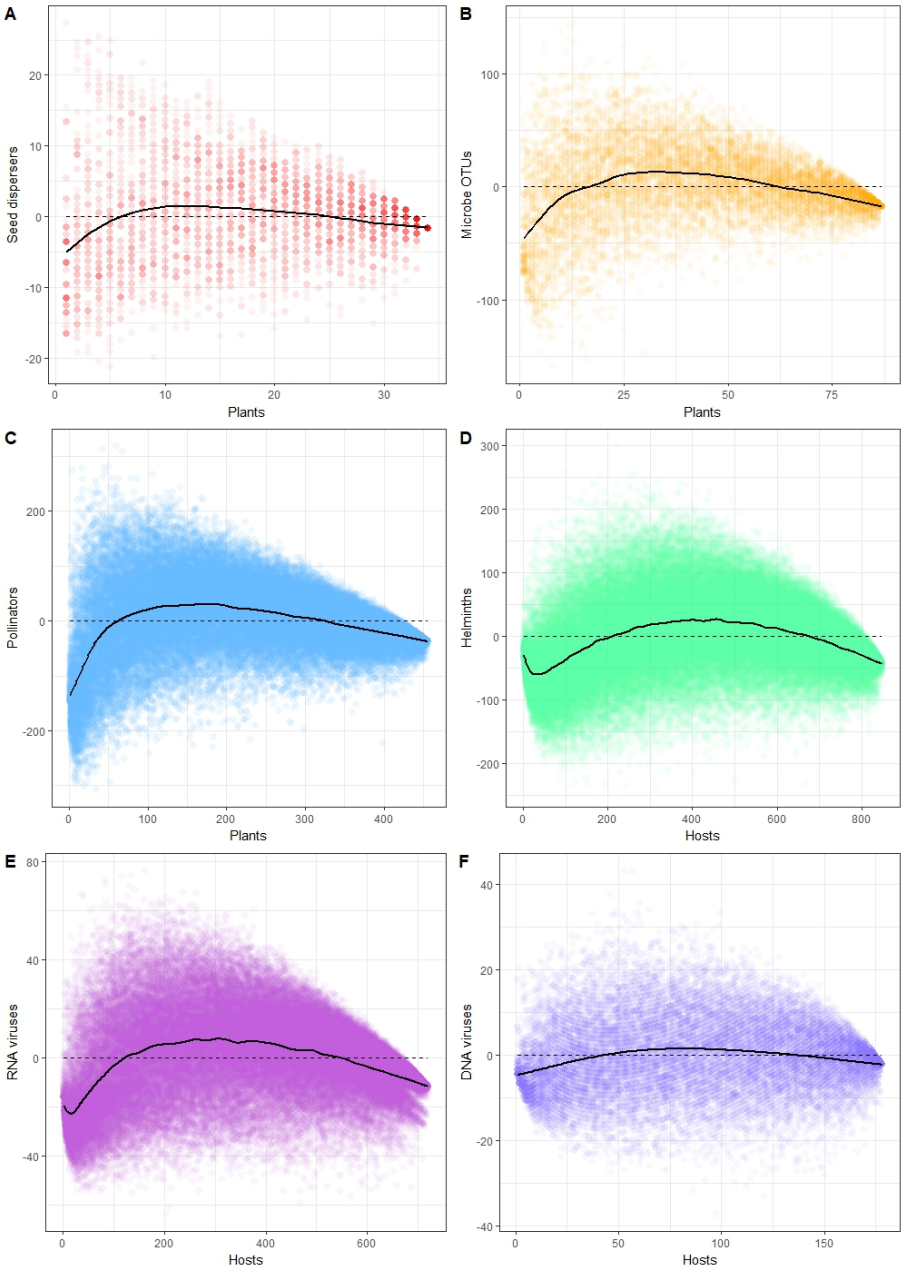


Figure S4: Residuals of a power law fit with `nls` to each of the six power law curves in Figure S2, with a smoothed spline fit through each using the default `smooth.spline` procedure in `R`. These curves are all fit with 100 iterations generated in the subsampling procedure.

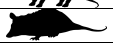
		Zoonotic Virus Network			Non-zoonotic virus network		
Order icon	Order name	Degree	Average Node Strength	Number of samples	Degree	Average Node Strength	Number of samples
	Carnivora	13	0.175	145	5	0.050	106
	Cetartiodactyla	13	0.164	211	7	0.033	344
	Chiroptera	13	0.138	420	2	0.023	89
	Cingulata	10	0.13	8	N/A	N/A	N/A
	Didelphimorphia	11	0.148	25	N/A	N/A	N/A
	Diprotodontia	10	0.065	12	N/A	N/A	N/A
	Eulipotyphla	11	0.053	9	N/A	N/A	N/A
	Human	13	0.166	188	N/A	N/A	N/A
	Lagomorpha	13	0.127	21	2	0.028	18
	Peramelemorphia	N/A	N/A	N/A	1	0.010	2
	Perissodactyla	13	0.163	43	5	0.061	47
	Pilosa	11	0.134	17	N/A	N/A	N/A
	Primates	13	0.113	250	3	0.026	82
	Proboscidea	11	0.058	5	4	0.040	4
	Rodentia	13	0.160	458	7	0.026	208
	Scandentia	N/A	N/A	N/A	N/A	N/A	N/A

Table S1: Network properties of orders and key to images from Figure 2.

Dataset	Power	Linear	Log
Pollinator	**515327.74**	570091.01	528984.14
Seed dispersal	**22740.90**	24271.36	23159.80
Plant-microbe	**86334.48**	93360.59	88638.36
Host-helminth	**941678.37**	1039936.03	1165296.40
Mammal-RNA virus	**621457.06**	694165.89	672021.56
Mammal-DNA virus	**128593.21**	136563.00	151715.53

Table S2: AIC values for power law, linear, and log forms tested for six datasets (shown in Figure S2).

Dataset	Estimate	Lower	Upper
Pollinator	0.3750	0.3740	0.3760
Seed dispersal	0.4651	0.4583	0.4719
Plant-microbe	0.4540	0.4508	0.4572
Host-helminth	0.6791	0.6784	0.6797
Mammal-RNA virus	0.4615	0.4605	0.4625
Mammal-DNA virus	0.6659	0.6632	0.6687

Table S3: Point estimates and 95% confidence intervals for z generated from subsampling 100 iterations.

Model name	Model
Classic power law	$A = b_0 H^{b_1}$
Quadratic power law	$A = \exp(b_0 + b_1 \log H + b_2 (\log H)^2)$
Extended power model 1 (EPM1)	$A = b_0 H^{b_1 H^{-b_2}}$
Extended power model 2 (EPM2)	$A = b_0 H^{b_1 - b_2/H}$
Persistence function 1 (P1)	$A = b_0 H^{b_1} \texttt{exp}(-b_2 H)$
Persistence function 2 (P2)	$A = b_0 H^{b_1} \texttt{exp}(-b_2/H)$

Table S4: Six candidate models.

Model	1	2	3	4	5	6
PL	22741	86334	515328	941678	128593	621457
QPL	22568	**84990**	**504340**	**925529**	127979	**606716**
EPM1	22565	—	—	925594	**127976**	606750
EPM2	22664	86182	506064	926903	128049	608016
P1	22599	85145	506291	926331	128055	608016
P2	**22559**	—	—	928453	128032	608894

Table S5: Model selection exercise, using the models presented in Table S4 with 100 iterations each. AICs are calculated for models fit to six network datasets: 1) plant-seed disperser, 2) plant-mycorrhizal OTU, 3) plant-pollinator, 4) host-helminth, 5) host-DNA virus, and 6) host-RNA virus. Dashed values indicate a model did not converge. Bolded values are the lowest AIC for a given curve

	DNA virus richness		RNA virus richness	
Model	Raw	Corrected	Raw	Corrected
PL	1,608.9	26,375	895.6	14,681
QPL	749.6	12,289	519.4	8,515
EPM1	828.1	13,575	568.8	9,325
EPM2	1,186.9	19,457	710.4	11,646
P1	4.8	79	116.8	1,915
P2	1,335.9	21,900	751.0	12,311

Table S6: Extrapolated predictions from all seven candidate model runs presented in Table S4.

DNA viruses		Estimate	95% CI
100% method	Raw estimate	1,611	(1,593—1,631)
	Sampling correction	6,312	(6,117—9,010)
50% method	Raw estimate	2,290	(2,243—2,339)
	Sampling correction	8,970	(8,612—12,922)

RNA viruses		Estimate	95% CI
100% method	Raw estimate	893	(889—897)
	Sampling correction	12,381	(10,132—27,571)
50% method	Raw estimate	1,126	(1,118—1,1135)
	Sampling correction	15,624	(12,746—34,902)

Table S7: Separate DNA and RNA rates of sampling