PNAS
www.pnas.org

SI Appendix

True scale-free networks hidden by finite size effects

Matteo Serafino, Giulio Cimini, Amos Maritan, Andrea Rinaldo, Samir Suweis, Jayanth R. Banavar, Guido Caldarelli

Corresponding Author name: Guido Caldarelli.
E-mail: guido.caldarelli@unive.it

This PDF file includes:

SI Appendix Text

Degree cross-over k_c versus maximum degree k_{max}

The maximum degree in a network is defined as the value of k for which the probability to finding a node with equal or higher degree is $1/N$. For a power law degree distribution,

$$\int_{k_{max}}^{\infty} k^{-\lambda} dk \sim \frac{1}{N} \quad [1]$$

from which it follows that $k_{max} \sim N^{1/(\lambda-1)}$ (1). The degree cross-over is instead the value of the degree for which the distribution has a crossover from a pure power law behavior to a faster decay (due to finite size effects). This is what defines k_c in the FSS ansatz. Using the same steps of eq. (1) for $p(k,N) = k^{-\lambda} f(k/k_c)$ we get

$$\begin{aligned} \int_{k_{max}}^{\infty} k^{-\lambda} f(k/k_c) dk &= k_c^{1-\lambda} \int_{\frac{k_{max}}{k_c}}^{\infty} x^{-\lambda} f(x) dx \\ &= k_c^{1-\lambda} F(k_{max}/k_c) \sim \frac{1}{N} \end{aligned} \quad [2]$$

and again we find $k_c \sim N^{1/(\lambda-1)}$. Hence despite the values of k_{max} and k_c are different, both of them have the same scaling behavior with N. However in the main paper we showed that $k_c \sim N^{1/\lambda}$, at stake with eq. (2). Moreover, Figure S2 shows that k_{max} also scales with $1/\lambda$ in our empirical data. Therefore k_{max} and k_c show the same dependence on N, but with an exponent different from the expected one. We believe this is due to the correlations inside the networks which alter the dependence on N (2). Another argument in favor of the same scaling of k_c and k_{max} is as follows. By definition, k_c is the value of k for which the argument of the scaling function becomes constant, whence $k_c \sim N^{-d}$. Concerning k_{max}, since the cumulate of the distribution of maxima is

$$P^{>}_{max}(k,N) = N k^{-\gamma} f(kN^d) \quad [3]$$

for not too small k, we have that the characteristic maximum, defined as $k^*_{max} \equiv \left\langle k^i \right\rangle_{max} / \left\langle k^{i-1} \right\rangle_{max}$, where $i > \gamma + 1$ and $\langle \cdot \rangle_{max}$ is the average with the distribution $-\frac{d}{dk} P^{>}_{max}(k,N)$, behaves as $k^*_{max} \sim N^{-d}$ for large N.

Finite-size scaling and system size

In order to apply finite-size scaling to network data we start from an empirical network of N nodes and then build smaller sub-networks through a sampling scheme (see below for more details). Given the values of N typical of empirical networks, the down-scaling of the system is performed linearly, whence the sub-networks of sizes $\frac{N}{4}$, $\frac{N}{2}$ and $\frac{3N}{4}$ considered in this study. However, when considering artificial network data we are free to down-scale the system even by orders of magnitude. This is shown in Figure S3 in the case of a Barabási-Albert model with $N = 10^5$. Notably, parameters estimated on this systems are perfectly in line with those estimated with the linear sub-sampling scheme (see main paper and Figure S6).

Node sampling method

The down-scaling of a network, namely how to derive a representative sample of the original network, is still an open problem nowadays. There are two main approaches in the literature to deal with this issue: sampling (3–5) and renormalization (6–9). Despite their simplicity, sampling procedures do not necessary preserve the properties of the original network. On the contrary, renormalization (i.e., coarse grain) procedures are based on the conservation of these properties during the down-scaling; however these methods lack of generality by requiring ad hoc assumptions. In this paper we decide to use the most general, although not necessary the most accurate, scheme: random node sampling. Here we provide a detailed discussion on the degree properties of our sampled networks. To build a sub-network instance, a set of n nodes (with $n \leq N$) are selected uniformly at random. The sampled network is the network induced by these nodes and by all the links among them. See Figure S4 for an sample illustration. Figure S5 reports the degree distributions of the (sub-)network in the case of a Barabási-Albert model with $N = 10^4$ and various densities ($\langle k \rangle = 1$, 5, 14 for the first, second and third column respectively). As in the main paper, we chose $n \in \{\frac{N}{4}, \frac{N}{2}, \frac{3N}{4}, N\}$ as possible size of the sub-network. However, different sizes of the network may be chosen. Indeed, as we show in Figure S3, results do not change by changing the sub-networks dimension. In the first row of the figure (panels $a-c$) we report the cumulative degree distributions for a single instance of a sub-sampled network. In agreement with (4), for nodes with low connectivity and for $n \ll N$ the random node sampling does not preserve the shape of the power law degree distribution. However these deviations are reduced by increasing the average degree of the network (i.e., by moving from the left to the right panel). In the second row of the figure (panels $d-f$) we report the cumulative degree distributions obtained by averaging over 100 instances of the sub-sampling. As expected, fluctuations in the degree are drastically reduced. Notably, in the region between k_{min} and k_{max} where the scaling assumption holds, the original degree distribution is preserved. This is particularly evident when $\langle k \rangle > 1$. However, because the down-scaling changes the mean degree of the networks, these curves do not collapse on top of each other as one would expect. This issue also affects renormalization procedures (6, 9), because one should compare *rescaled* quantities. To do that there are two options. The first is to work with rescaled degrees $k/\langle k \rangle$ instead of actual degrees (9). As shown in the third row of the figure (panels $g-i$), by doing this all the distributions follow a single

master curve for $k \geq k_{min}$. The other possibility is to adjust the average degree without modifying the statistical properties of the network (6), for instance by taking the part of the distribution beyond k_{min} (estimated by maximum-likelihood power law fitting on the whole network (10, 11)) and renormalizing it such that $\sum_{k=k_{min}}^{k_{max}} P(k) = 1$. By doing this, as shown in the fourth row of the figure (panels $j - l$), the distributions for different n still collapse on the top of each other (without the need to rescale by the mean degree), apart for the region near k_{max} where finite size effects step in. Notably, the mean degree does not change as an effect of the cut of the distribution below k_{min}. Indeed, let us denote by n^* the number of nodes with $k \geq k_{min}$, and by e^* the number of links connecting these n^* nodes among themselves (l links) plus the number of links connecting these n^* nodes with the other nodes in the network with $k < k_{min}$ (s stubs). For scale-free network with $\lambda > 2$, the mean degree is finite and since $\langle k \rangle = 2E/N \sim$ constant we have $E \sim N$. We measured n^* versus e^* for 20 equally spaced values of n. A fitting of $\ln e^*$ vs $\ln n^*$ yields a slope equal to 1.02 ± 0.10, in agreement with the expected value of one obtained when the mean degree is preserved during the down-scaling of the network. In conclusion, as long as the considered network has an average degree much bigger than one (as in our case, see the left panel of Figure S6), the sub-sampling procedure preserves pretty well the degree distribution of the mother network. However, we remark that our random sampling procedure cannot substitute more rigorous renormalization methodologies, since it does not allow us to have a unique representation of the reduced network. Finally let us notice that the reason why we use the cut procedure is because we are interested in computing the quality of the collapse only in the region where the scaling hypotheses holds, i.e., when $k \geq k_{min}$.

Choice of the average degree

In the main paper we study the benchmark Barabási-Albert model with $m \equiv \langle k \rangle = 14$. We choose this value in accordance with the mean of the $\langle k \rangle$ in the empirical network of our dataset (see panel a of Figure S6). Other choices of m are however possible. Panels b and c of Figure S6 show the scaling analysis with N and E, respectively, for a Barabási-Albert model with $m = 5$. The power law exponent estimated by means of the KS test is $\Gamma = 1.847 \pm 0.039$. In the first case the quality of the collapse is $S = 0.26$. $\gamma = 1.807 \pm 0.108$ while $d = -0.378 \pm 0.133$. In the second case we have $S = 0.25$, $\gamma = 1.807 \pm 0.113$ and $d_E = -0.361 \pm 0.119$. The network is classified as strongly scale-free. Note that the quality of the collapse S does not depend on the average degree of the network (both values $m = 5$ and $m = 14$ give a close result for S). Moreover, the value of the estimated exponent γ is different from 2 in both cases. This is not a consequence of using a large m, because $\gamma = 2$ holds in the asymptotic limit $N \to \infty$ while the number of nodes in our networks is always finite.

Clustering in the $\gamma - S$ plane

Panel a of Figure S7 shows the scatter plot of γ versus S, with each point representing an empirical network. As mentioned in the main text, we observe no clusterization of networks amenable to categories. This happens however because we built the dataset balancing the various network categories (see details in the Methods section). Networks that are very similar may instead cluster together, as shown in panel b of the same figure where we report all the 109 KEGG protein interaction networks that can be found on ICON.

Binning errors

As explained in the main paper, in order compute the quality of the collapse S we need to assign an error $\sigma(k)$ to each binned value of $p(k)$ in each (sub-)network. When degrees are independent and not correlated, then the counts in each bin k should closely follow a Poisson distribution. In this case a good estimate for the error in each bin is simply the square root of the counts: if $p(k) = n_k/n$ is the fraction of nodes with degree k in a set of n nodes, then the error is $\sigma(k) = \sqrt{n_k}/n$. However in real networks degrees are typically correlated, therefore this approach could lead to uncontrolled biases. An alternative way that accounts for correlations is to compute the errors by means of a bootstrapping approach. Given a degree sequence of length n we uniformly at random extract $\tilde{n} \leq n$ values (extractions with replacement). By repeating this procedure an arbitrary number of times and by computing the standard deviation of the counts in each bin among all the realizations we obtain an unbiased estimation of the error.

Panel a of Figure S8 shows the scatter plot of the errors obtained using the two described approaches, in the case of a numerical realizations of the Barabási-Albert model with $m = 14$ and $N = 10^4$. In particular we consider the sub-network of size $3N/4$, whose degree distribution is obtained by averaging over multiple (100) sub-samples. Hence for the bootstrap case we have a degree sequence of total length $n = 100 \times 3N/4$, from which we generate 10^4 bootstrapped sequences of length $\tilde{n}$. We use $\tilde{n} = n$ (remember that we allow for multiple extractions of the same entry). The two errors are very similar (linear fitting returns a slope close to one). Hence the two approaches are almost equivalent – for computational reasons we used the Poissonian approach. Panel b of the figure further shows that the higher difference between the two errors is obtained especially for high values of k. At last in panel c we show that the average ratio of the Poissonian and bootstrapped errors is inversely proportional to $\tilde{n}$, but stays in the range $[1, 2]$ until very low values of $\tilde{n} \sim n/10$.

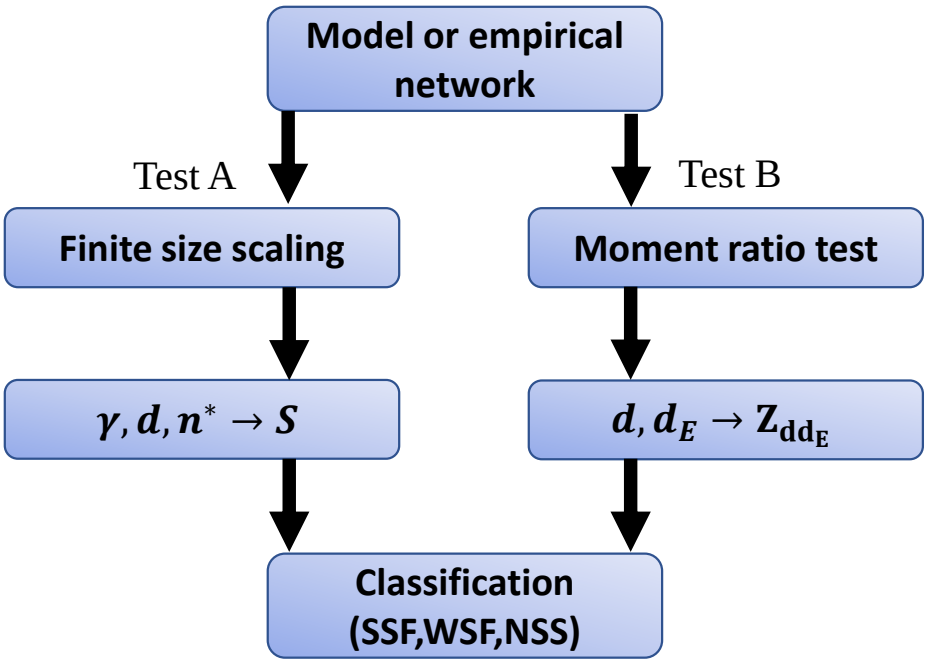

Fig. S1. Schematic flow of the analysis. For a given network we independently perform two tests: the finite size scaling (A) and the moment ratio test (B). The respective statistical outputs (the quality of the collapse S and the Z-score between d and d_E) are combined to obtain a classification of the networks as **SSF**,**WSF** or **NSF**.

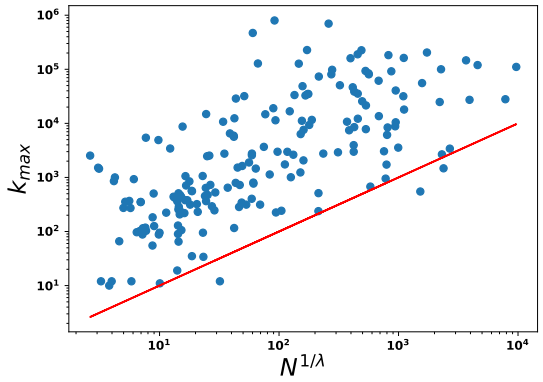

Fig. S2. Empirical relation of k_{max} versus $N^{1/\lambda} = N^{1/(\Gamma+1)}$ for the real networks in our dataset. Here Γ is the exponent estimated via maximum-likelihood fitting (10). The red line represents the identity and serves as a guide for the eye.

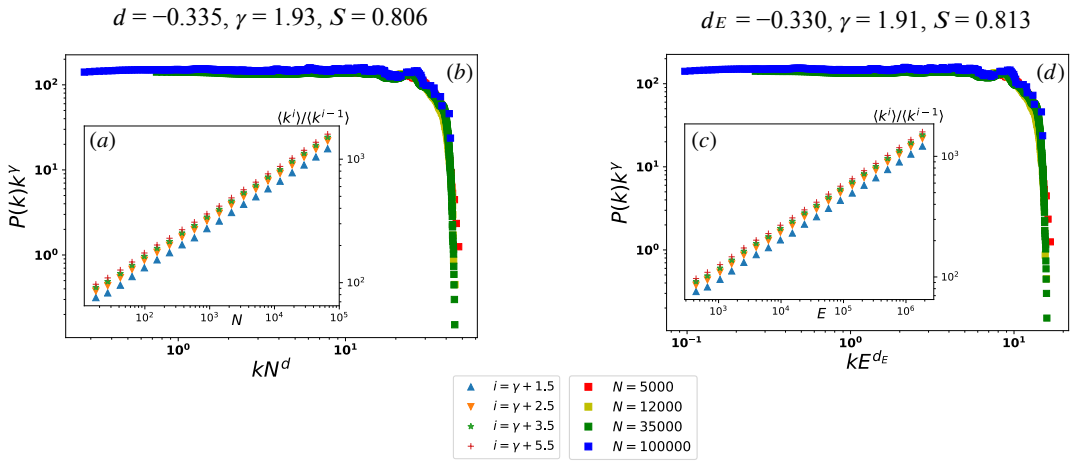

Fig. S3. Scaling analysis on a numerical realization of the Barabási-Albert model with $N = 10^5$ and $k_{min} = 14$. Panels (a), (b): the scaling analysis with N yields $d = -0.335 \pm 0.024$, $\gamma = 1.93 \pm 0.03$ and $S = 0.81$. Panels (c), (d): the scaling analysis with E yields $d_E = -0.330 \pm 0.024$, $\gamma = 1.91 \pm 0.03$ and $S = 0.81$.

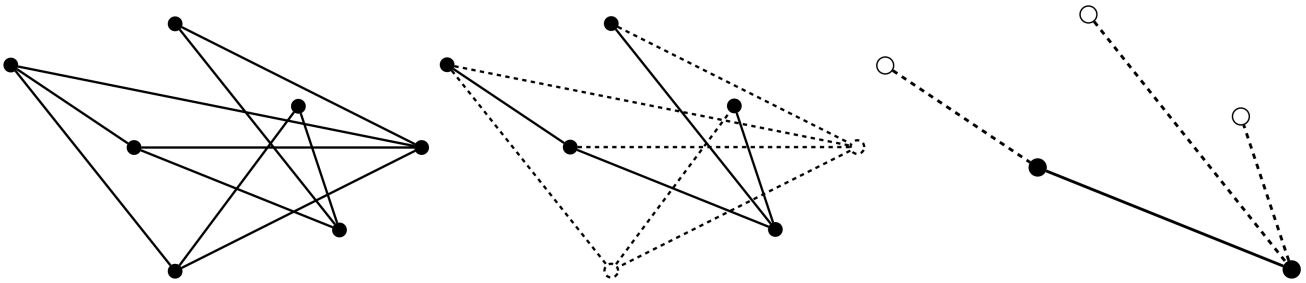

Fig. S4. Illustration of the sub-sampling scheme. The original network is reported on the left. In the middle we report the sub-network induced by 5 random selected nodes, with removed links/nodes represented with dashed lines. On the right an example of the counting procedure after the cut at k_{min}. If $k_{min} = 2$, then $N^* = N_{k \geq k_{min}} = 2$ (filled circles) and $E^* = l + s = 1 + 3 = 4$, with l number of links (solid lines) and s number of stubs (dashed line).

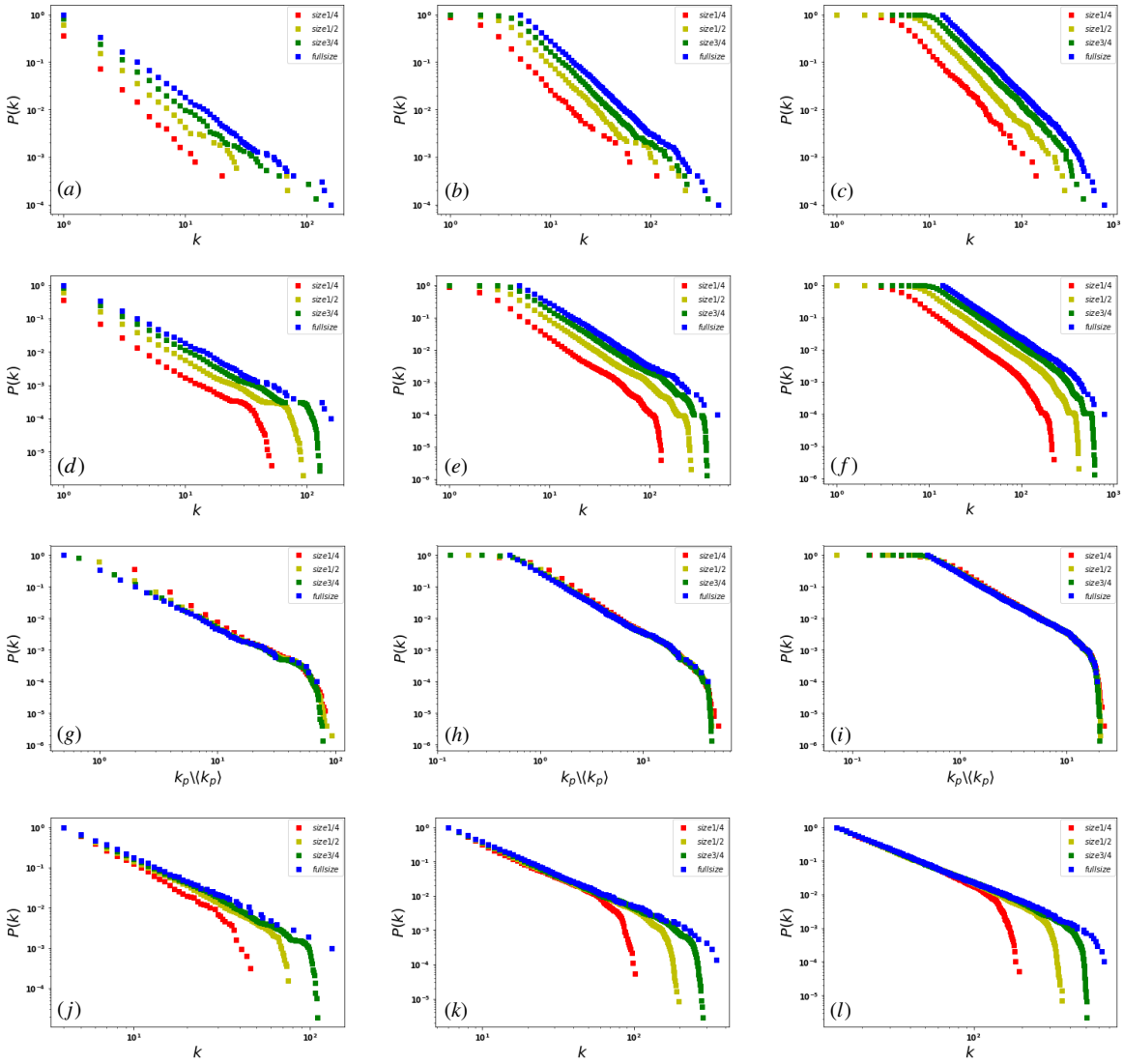

Fig. S5. Analysis on three different realizations of a Barabási-Albert model with $N = 10^4$ and $\langle k \rangle = 1, 5, 14$, from the left to the right. On the top the degree distributions extracted using single sub-sample. In the second row the results for multiple sampling for each dimension n of the (sub-)network. The third row reports the same results of the second row, but with rescaled degree sequences. On the bottom the result with multiple sampling for each value of n after the cut and the normalization.

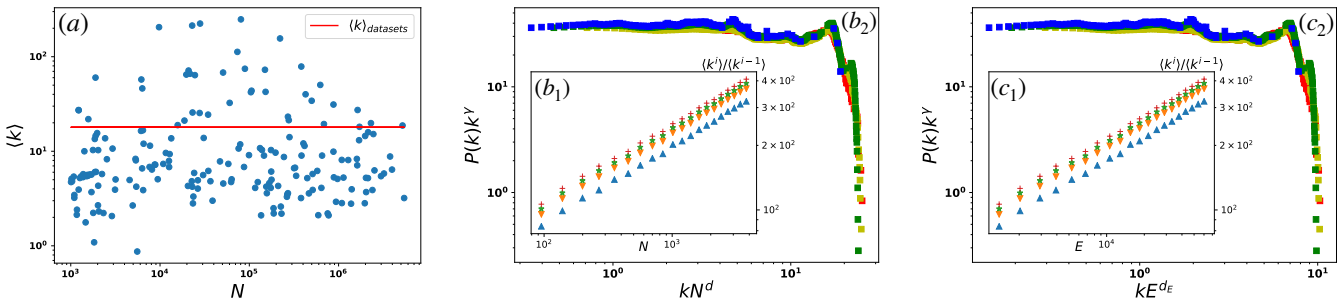

Fig. S6. (a): Mean degree $\langle k \rangle$ as a function of the number of nodes N. Each point corresponds to an empirical network in our dataset, while the red line is the value $\langle k \rangle = 14$ we use in the benchmark Barabási-Albert model. The other two panels (b, c): report the scaling analysis with N and E in the case of a Barabási-Albert network with $m \equiv \langle k \rangle = 5$.

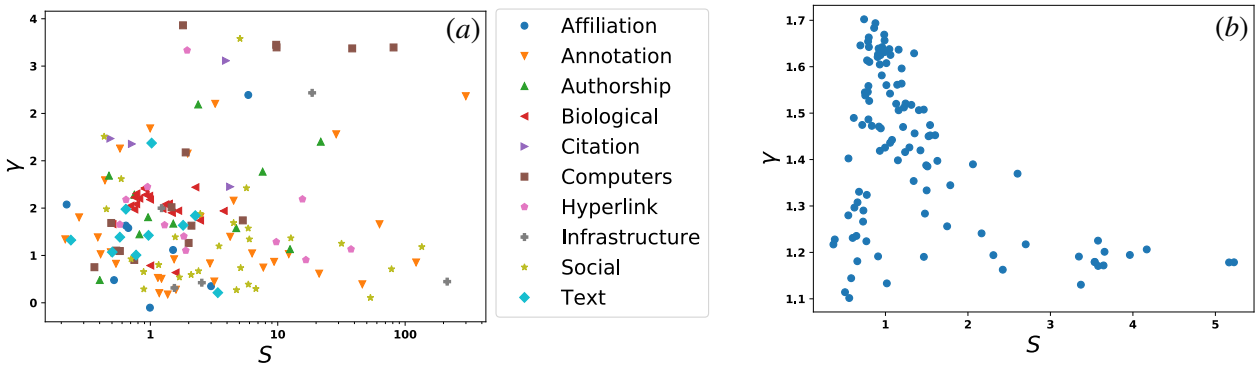

Fig. S7. Relation between the power law exponent γ and the quality of the collapse S resulting from the finite size scaling analysis with N. Each point represents an empirical network. Panel a reports results for the whole dataset at our disposal (all the networks for which S is defined), whereas, panel b focuses only on metabolic networks of various species from KEGG (12) – which are mostly not included in panel a.

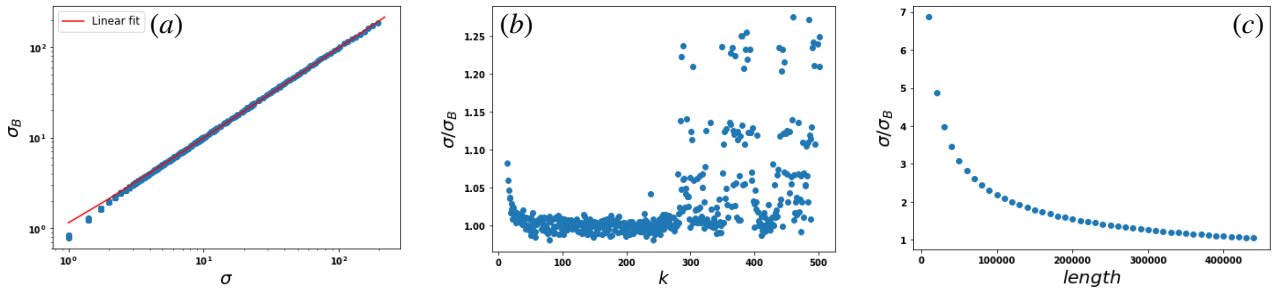

Fig. S8. (a) Scatter plot of σ_B (the error computed with the bootstrap procedure) versus σ (the Poissonian error obtained as the square root of the counts), in the case of a Barabási-Albert model with $m = 14$ and $N = 10^4$. We use $\tilde{n} = n$. The linear fit (shown as a red line) has a slope equal to 0.966 ± 0.001. (b) Ratio σ/σ_B as a function of the binned degree k for $\tilde{n} = n$. (c) Mean ratio σ/σ_B as a function of the length $\tilde{n}$ of the bootstrapped sequence.

References

1. SN Dorogovtsev, JFF Mendes, AN Samukhin, Structure of growing networks with preferential linking. *Phys. Rev. Lett.* **85**, 4633–4636 (2000).
2. M Boguñá, R Pastor-Satorras, A Vespignani, Cut-offs and finite size effects in scale-free networks. *The Eur. Phys. J. B* **38**, 205–209 (2004).
3. J Leskovec, C Faloutsos, Sampling from large graphs in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 06. (Association for Computing Machinery, New York, NY, USA), pp. 631–636 (2006).
4. MPH Stumpf, C Wiuf, RM May, Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proc. Natl. Acad. Sci.* **102**, 4221–4224 (2005).
5. SH Lee, PJ Kim, H Jeong, Statistical properties of sampled networks. *Phys. Rev. E* **73**, 016102 (2006).
6. G García-Pérez, M Boguñá, MA Serrano, Multiscale unfolding of real networks by geometric renormalization. *Nat. Phys.* **14**, 583–589 (2018).
7. C Song, S Havlin, HA Makse, Self-similarity of complex networks. *Nature* **433**, 392 (2005).
8. D Gfeller, P De Los Rios, Spectral coarse graining of complex networks. *Phys. Rev. Lett.* **99**, 038701 (2007).
9. MA Serrano, D Krioukov, M Boguñá, Self-similarity of complex networks and hidden metric spaces. *Phys. Rev. Lett.* **100**, 078701 (2008).
10. A Clauset, CR Shalizi, ME Newman, Power-law distributions in empirical data. *SIAM review* **51**, 661–703 (2009).
11. J Alstott, E Bullmore, D Plenz, powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE* **9**, e85777 (2014).
12. M Huss, P Holme, Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. *IET Syst. Biol.* **1**, 280–285 (2007).