Measure for characterizing heavy-tailed networks

Sam A. Hill*
Department of Physics, Adrian College, Adrian, Michigan 43606, USA

(Received 19 December 2019; accepted 11 May 2021; published 30 June 2021)

The phrase "scale-free network" has become controversial in recent years, as network scientists debate what it means for a finite degree sequence to fit a power-law distribution. In practical terms, however, most network scientists use the phrase to indicate that a network has hubs, and so it would be useful to be able to talk about such networks without reference to power laws at all. This paper presents the Cooke-Nieboer index (CNI), a nonasymptotic measure of the heavy-tailedness of a network's empirical degree distribution which does not presume a power-law form. The CNI is easy to calculate and is able to distinguish between synthetic networks with power-law, exponential, and symmetric degree distributions. It serves as a complementary measure to the traditional tail-index estimators and reflects certain properties in real-life networks better than the estimators do.

DOI: 10.1103/PhysRevResearch.3.023257

I. MOTIVATION

A phase change in network science research occurred at the end of the last century, with the discovery that the relationships in many real-life systems have properties which can not be captured by Erdős-Rényi random graphs [1]. One important such property is the existence of hubs [2]: nodes with large degrees, much larger than an Erdős-Rényi graph of the same size and average degree would possess. Hub-dominant networks are everywhere [3], from the structure of the Internet, to metabolic networks, to friendships and followers both online and in real life, and in many other examples. Hub-dominated networks are usually referred to as "scale-free networks" in the literature, which implies that their degree distribution $P(x)$ corresponds in some way to a power law:

$$P(x) \sim x^{-\alpha-1}, \quad \alpha > 0. \tag{1}$$

The meaning of "$\sim$" in the equation above is very flexible, however, and as Broido and Clauset [4] point out, there are a variety of opinions about what "scale-free" really means. Some authors [2–5] require that the degree distribution of a "scale-free" network, or at least a portion of that distribution, must follow a strict power law. Others are more lenient, requiring that the degree distribution be regularly-varying [6] or heavy-tailed, that the distribution be "well-approximated" by a power law [5], or even that the distribution "looks linear" on a log-log plot (as discussed in Ref. [7]). Some use the term "scale-free" to describe aspects of a network which are unrelated to its degree distribution, such as the self-similarity of its subgraphs [8,9].

The recent debate [4,10,11] over this term could be dismissed as merely a semantic one, but there are a few problems with the widespread use of the term "scale-free." First, real-world networks are always finite: their degree sequences end, their variance is finite, and by definition (see Eq. (2) below; also Ref. [12]) they cannot be heavy-tailed, let alone scale-free. Researchers work around this by imagining that real-world networks are subsamples of some underlying infinite distribution [6] or generated via some well-defined process which would create a power-law distribution in the infinite-size limit [11]. This assumes, however, that some such distribution or process exists.

Second, sometimes the distinction between true power-law and mere heavy-tailed networks is important: for example, the proof [13] that certain scale-free networks have no epidemic threshold depends on the infinite variance of a power-law degree distribution with $\alpha \leqslant 2$; hub-dominated networks with finite variance may not share this property.

A third problem with this approach is that it encourages researchers to use power-law measures to characterize hub-dominant networks. One common way to characterize the "scale-freeness" of a network is to fit a portion of the degree sequence (usually the "tail" where the degrees are highest) to a power law [Eq. (1)]; the fit parameter α is often known as the *tail index* of the network [4,14–17]. But if the finite network has only a few high-degree nodes, this asymptotic measure becomes sensitive to the properties of a very small portion of the network.

When we get caught up in the details of power-law fits and statistical tests, we overlook the fact that when most network scientists describe a network as "scale-free," they are referring to the existence of hubs. The presence of hubs in a network gives them distinctive properties: networks with hubs have shorter path lengths [5], allow for the efficient spread of information (or viruses) [13], and are robust against random failures [18] but more susceptible to targeted attacks [19].

*shill@adrian.edu

Therefore it may be useful to think about the existence of hubs in a different, noninferential way.

Therefore we here present a new measure: the *Cooke-Nieboer index* (CNI), which attempts to quantify the presence of hubs in a network. This measure, adapted from the "obesity index" in Ref. [14], does not presume that the distribution is scale-free, nor is the measure asymptotic: it is applied to the entire degree sequence and not just to its tail. In this paper we will define the CNI and investigate its behavior for several standard mathematical distributions and synthetic networks, classifying them into "high," "low," and "negative" CNI categories. We will also apply our measure to real-world networks, comparing the CNI with the alternate classification schemes found in Refs. [4,6]. We will give several examples where the CNI may give a more accurate representation than the tail index of the underlying network.

II. DEFINITION

A. The obesity index

In the probability literature [12], the probability density function (PDF) $f(x)$ of a distribution is said to be *heavy-tailed* if

$$\int_{-\infty}^{\infty} e^{\lambda x} f(x)\,dx = \infty \quad \text{for all } \lambda > 0. \tag{2}$$

This implies that the PDF decays more slowly than any exponential. Most heavy-tailed distributions of interest fall into a subcategory known as the *subexponential* distributions, defined as follows [20]: if $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables chosen from a subexponential distribution, then

$$\lim_{x\to\infty} \frac{P(X_1 + \cdots + X_n > x)}{P(\max(X_1, \ldots, X_n) > x)} = 1, \quad \text{for all } n \geqslant 1. \tag{3}$$

In other words, the sum of the random variables is likely to be large if and only if their maximum is likely to be large. This is the *principle of a single big jump* [12]. (For example, if the cost of cleaning up from natural disasters follows a subexponential distribution, then the total cost of cleanup in any given year is going to be roughly equal to the total cost of the largest disaster that year.) Power-law and regular-varying distributions [6] are examples of subexponential distributions.

To characterize the "subexponentiality" of a distribution X, Cooke and Nieboer [14] suggest a measure known as the *obesity index*.

Definition. Select a quadruple (i.e. a set of four numbers) of i.i.d. random values from the distribution X, and label them in ascending order, so that $X_1 \leqslant X_2 \leqslant X_3 \leqslant X_4$. Then the obesity index is the probability

$$\mathrm{Ob}(X) \equiv P(X_1 + X_4 > X_2 + X_3). \tag{4}$$

If the distribution is symmetric, then the quantities $X_4 + X_1$ and $X_2 + X_3$ are equally likely to be larger, and so the distribution's obesity index is one-half [14]. For a subexponential distribution, on the other hand, Eq. (3) becomes

$$\lim_{x\to\infty} \frac{P(X_1 + X_2 + X_3 + X_4 > x)}{P(X_4 > x)} = 1, \tag{5}$$

which means that X_4 has a high probability of being larger than the sum of the other three variables, in which case $X_1 + X_4$ will normally be greater than $X_2 + X_3$, and the probability in Eq. (4) will be much greater than one-half.

The obesity index is a probability, and so ranges from zero to one. Like skewness and kurtosis, it is independent of offset and positive scaling of the distribution: i.e.,

$$\mathrm{Ob}(aX + b) = \mathrm{Ob}(X), \quad a \in \mathbb{R}^+, \quad b \in \mathbb{R}. \tag{6}$$

Multiplying the distribution by a negative number reverses the inequality in Eq. (4), however, so that

$$\mathrm{Ob}(b - aX) = 1 - \mathrm{Ob}(X), \quad a \in \mathbb{R}^+, \quad b \in \mathbb{R}. \tag{7}$$

B. The Cooke-Nieboer index

The obesity index is specifically designed for continuous distributions, and so needs to be modified before we can apply it to sequences of integers. We here propose a variation of the Obesity Index, which we call the *Cooke-Nieboer index* $\Theta(X)$. For a given distribution X, we define the CNI in the following way.

Definition. Let $X_1, \ldots, X_4$ be four i.i.d. random numbers chosen from a particular distribution X. We define the *Cooke-Nieboer index* of the distribution to be

$$\Theta(X) \equiv E\left\{\mathrm{sgn}\left(\frac{1}{2}(\max X_i + \min X_i) - \frac{1}{4}\sum_i X_i\right)\right\}, \tag{8}$$

where $E\{\cdot\}$ signifies the expectation value and $\mathrm{sgn}(x)$ is the signum function

$$\mathrm{sgn}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}. \tag{9}$$

For later convenience, we define

$$\Phi(X) \equiv \tfrac{1}{2}(\max X_i + \min X_i) - \langle X_i \rangle \tag{10}$$

so that $\Theta(X) = E\{\mathrm{sgn}(\Phi(X))\}$.

The Cooke-Nieboer index differs from the obesity index in three ways: (i) it properly handles discrete distributions by accounting for the finite probability that $X_1 + X_4 = X_2 + X_3$; (ii) it is rescaled so that it ranges from -1 to 1, so that for symmetric distributions, $\Theta = 0$; and (iii) it avoids the term "obesity", which may cause confusion in applications of network science to health issues. For a continuous distribution X, the two measures are simply related:

$$\Theta(X) = 2\,\mathrm{Ob}(X) - 1. \tag{11}$$

One may interpret Eq. (10) in the following manner: the first term $\frac{1}{2}(\max X_i + \min X_i)$ is the halfway point between the largest and smallest values, and could be thought of as the "geometric center" of the quadruple (Fig. 1), while the second term $\langle X_i \rangle$ is of course the mean. When one of the values is much larger than the others, as is common for heavy-tailed distributions, it will pull the geometric center to the right of the mean, and so $\Phi(X)$ will be positive. (This makes the CNI a type of skewness measure for the distribution.) A negative CNI would occur, conversely, when there are many more large values in the distribution and only a few small ones.

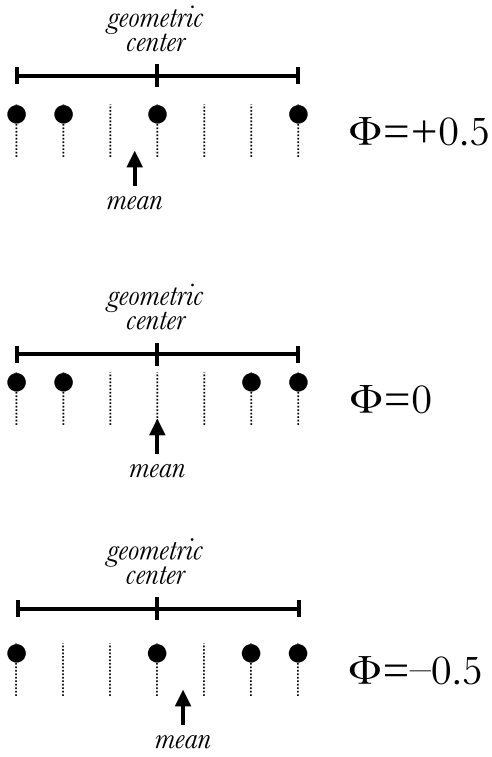

FIG. 1. Three examples of quadruples and their corresponding values of Φ as calculated by Eq. (10). The "geometric center" is the halfway point between the minimum and maximum values, and Φ measures whether the geometric center is to the right ($\Phi > 0$) or the left ($\Phi < 0$) of the mean. Notice that $\Phi = 0$ for the distribution that is symmetric about its geometric center.

An integral expression for the CNI can be derived from a similar integral in Ref. [14], but for most cases cannot be expressed in closed form. Instead, one may find the CNI of a continuous distribution by calculating Φ multiple times until reaching some desired standard error $\sigma_{\bar{x}}$, using code such as that found in Fig. 2. Figure 3 shows that the CNI calculated this way is normally distributed, with a standard

```
import numpy as np
from random import choices
def cni(degrees,maxerr=1e-3):
    vals=[]
    T=0
    while True:
      four=choices(degrees,k=4)
      phi=max(four)+min(four)-0.5*sum(four)
      vals+=[np.sign(phi)]
      T+=1
      sterr=np.std(vals)/np.sqrt(T)
      if(T>20 and sterr<maxerr):
        return np.mean(vals)
```

FIG. 2. Sample PYTHON code for calculating the CNI, given a list `degrees` of degrees of the network. The number 20 in the penultimate line is arbitrary and meant to prevent the code from stopping too soon. The code is written for demonstration purposes and is not particularly efficient; a more sophisticated version can be found in Appendix.

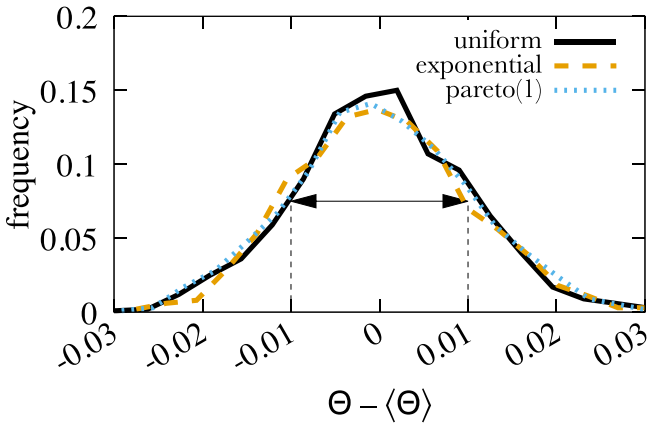

FIG. 3. The CNI of three different probability distributions—a uniform distribution between 0 and 1, an exponential distribution with $\lambda = 1$, and a Pareto distribution with $\alpha = 1$—was calculated one thousand times using a Monte Carlo algorithm such as Fig. 2, each time until reaching a standard error of $\sigma_{\bar{x}} = 0.01$. The figure shows the histogram of how the calculated Θ differs from the mean $\langle\Theta\rangle$ for that particular probability distribution. All three curves are localized and single-peaked with a standard deviation of 0.01, as expected.

deviation equal to $\sigma_{\bar{x}}$. The number of steps T required to reach a desired standard error is proportional to $\sigma_{\bar{x}}^{-2}$ (see Eq. (A2) of Appendix), with a coefficient depending on the type of distribution (Fig. 4).

III. DISTRIBUTION REGIMES

A. Bernoulli distribution

To understand how the CNI works, it is useful to consider the generalized *Bernoulli distribution*

$$X = \begin{cases} a & \text{with probability } p \\ b > a & \text{with probability } 1-p \end{cases}. \tag{12}$$

When we choose a quadruple from this distribution, and $0 \leqslant s \leqslant 4$ of the values are a, it is simple to show that Φ [Eq. (10)] is equal to zero if s is even, $\Phi < 0$ if $s = 1$, and

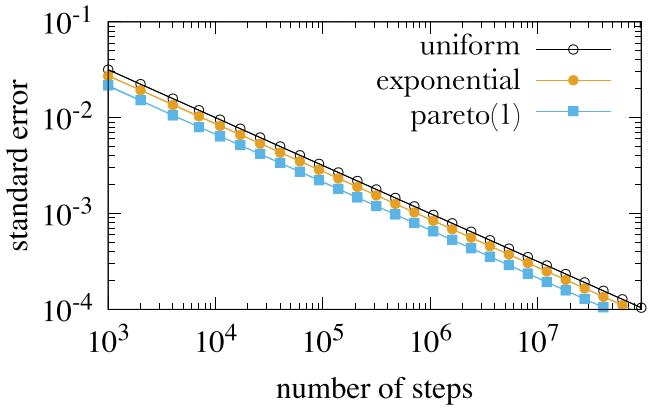

FIG. 4. For the same three distributions as in Fig. 3, this shows the number of steps T required to reach a particular standard error $\sigma_{\bar{x}}$, where a step is a single calculation of Φ [Eq. (10)]. All three curves closely obey the relationship $\sigma_{\bar{x}} \propto 1/\sqrt{T}$ after ten million steps.

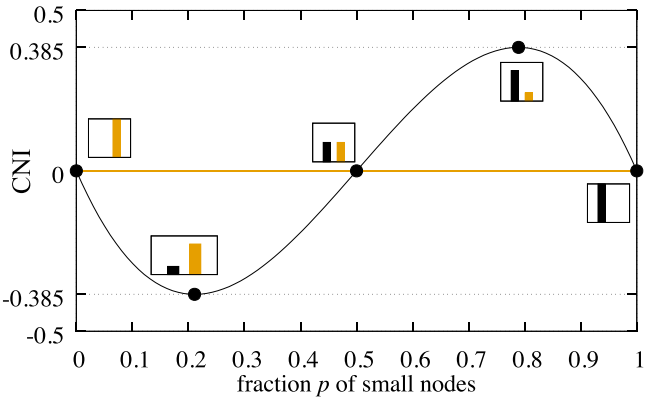

FIG. 5. The CNI of the Bernoulli distribution [Eq. (13)] as a function of p. The small boxes show the relative proportions of the two values ($X = a$ in black, $X = b > a$ in orange). The polynomial reaches extreme values of $\pm\frac{2\sqrt{3}}{9}$ at $p = \frac{1}{2} \pm \frac{\sqrt{3}}{6}$.

$\Phi > 0$ if $s = 3$. Thus we can calculate the CNI of this distribution precisely:

$$\begin{aligned}\Theta(p) &= \sum_{s=0}^{4} \binom{4}{s} p^s (1-p)^{4-s} \operatorname{sgn}(\Phi) \\ &= 4p^3(1-p) - 4p(1-p)^3 \\ &= 4p(1-p)(2p-1). \end{aligned} \tag{13}$$

Note that the result does not depend on the values a or b.

Figure 5 shows a graph of the polynomial in Eq. (13). Where the distribution is symmetric, at $p = 0$, 0.5, and 1, the CNI is zero; this follows from our discussion in Sec. II. When $p > 0.5$, there are more smaller values than larger values, and the CNI is positive; the CNI is negative when there are more larger values. The maximum CNI for a Bernoulli distribution is $\Theta = \frac{2\sqrt{3}}{9} \approx 0.385$ at $p = \frac{1}{2} + \frac{\sqrt{3}}{6} \approx 0.79$, which corresponds roughly to one large value out of every five.

B. Classification

In the previous section, we saw an example of the difference between distributions with positive and negative CNI, with the symmetric distributions ($\Theta = 0$) forming a boundary between the two regimes. We can further divide the positive regime into two classes using the exponential distribution $P(x) = \lambda e^{-\lambda x}$ as a second boundary. It is shown in Ref. [14] that the exponential distribution has an obesity index of 3/4 regardless of λ, and thus it has a $\Theta = 1/2$. Using the exponential distribution and the symmetric distribution values as boundaries, we propose to divide all distributions into one of three regimes.

(1) High-CNI distributions, with $\Theta > 0.5$. These are networks with a larger CNI than the exponential distribution. An important example are the power-law or Pareto distributions, whose CNIs (as shown in Fig. 6) range from $\Theta = 1$ for $\alpha = 0$ to $\Theta \to 0.5$ as $\alpha \to \infty$.

(2) Low-CNI distributions, with $0 \leqslant \Theta \leqslant 0.5$. This regime includes the symmetric distributions, the Gumbel distribution [14] $\exp(-e^{-x})$ (with $\Theta \approx 0.25$), and the Poisson distribution (as will be seen in Fig. 7).

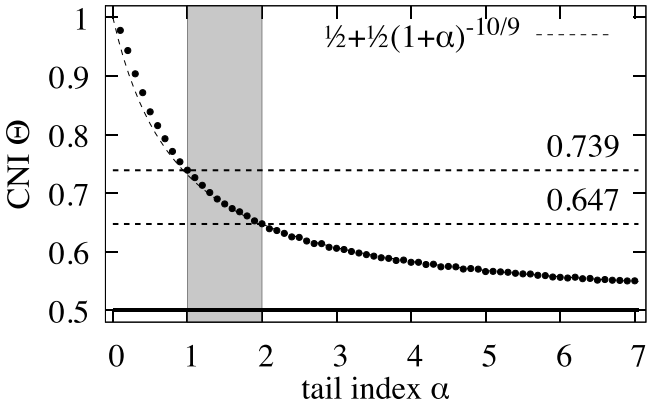

FIG. 6. The CNI of a power-law distribution $1/x^{\alpha+1}$ as a function of its tail index α. The CNI was calculated by selecting 1000 quadruples taken from the distribution; this process was repeated for 50 different sets of samples for each value of α. The standard deviation of these measurements are smaller than the height of the dots shown. The grey area highlights the region where most "scale-free" networks are found [4,21], between $\alpha = 1$ and $\alpha = 2$: Ref. [14] calculates the CNI at these values as $2\pi^2 - 19 = 0.739$ (for $\alpha = 2$) and $1185 - 120\pi^2 = 0.647$ (for $\alpha = 3$). There is no known closed form for this curve but it is close to the expression $\frac{1}{2} + \frac{1}{2}(1+\alpha)^{-10/9}$, which is shown as a dashed line.

(3) Negative-CNI distributions, with $\Theta < 0$. These are distributions which have many large values and fewer small values: distributions that are heavier on the right than on the left. We will see that many dense planar networks fall into this category.

IV. NETWORKS

The Cooke-Nieboer index can be extended to describe finite networks in a natural way. For an undirected, unweighted

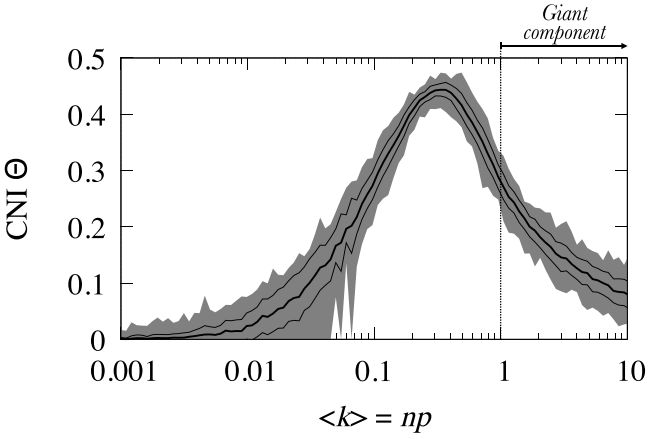

FIG. 7. The CNI Θ of Erdos-Renyi networks $G(n, p)$ of $n = 1000$ nodes with varying average degree $\langle k \rangle = np$. One hundred different networks were generated for each value of $\langle k \rangle$, and their CNIs were calculated to a standard error of 0.001. The thick central line shows the mean value of Θ; the two lines on either side show one standard deviation away from the mean. The shaded region shows the range of all values. Larger values of n (not shown) result in a similar trajectory but a smaller shaded region.

network G with degree sequence $\{k_1, k_2, \ldots, k_N\}$, we define $\Theta(G)$ to be the CNI of its degree sequence; that is, $\Theta(G) = E\{\mathrm{sgn}(\Phi)\}$, where

$$\Phi = \tfrac{1}{2}(\max k_i' + \min k_i') - \langle k_i' \rangle, \tag{14}$$

and k_i' are a set of four samples chosen from the degree sequence. For weighted networks, one can replace the degree k_i with the total weight of the edges connected to the node; note that there is no need for this to be an integer. We can also use the CNI to examine the distributions of other properties of networks, like the eigenvector or the betweenness distributions [3].

We can calculate the CNI by considering every combination of four elements of the degree sequence, allowing for duplicate selections to simplify the calculation:

$$\Theta = \frac{1}{N^4} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \mathrm{sgn}(\Phi(k_i, k_j, k_k, k_l)) \tag{15}$$

If the k_i are all integers less than or equal to some value M, and if p_a is the fraction of nodes with degree a, then we can write this in the more computationally efficient form

$$\Theta = \sum_{a=0}^{M} \sum_{b=0}^{M} \sum_{c=0}^{M} \sum_{d=0}^{M} p_a p_b p_c p_d \,\mathrm{sgn}(\Phi(a, b, c, d)), \tag{16}$$

which runs in $O(M^4)$ time as compared to $O(N^4)$ time.

The measure is not generally additive: the CNI of the union of two graphs G and H has no (known) simple relationship with $\Theta(G)$ and $\Theta(H)$, except when the networks are the same, in which case $\Theta(G \cup G) = \Theta(G)$ (due to the scaling independence Eq. (6) of the obesity index). From Eq. (7), it can be shown that the CNI of the complement $\bar{G}$ of a graph G (that is, a graph where two nodes are linked in $\bar{G}$ if and only if they are not linked in G) is

$$\Theta(\bar{G}) = -\Theta(G). \tag{17}$$

Networks with symmetric degree distributions, such as complete graphs and cycle graphs, have $\Theta = 0$.

A. Erdős-Rényi networks

Erdős-Rényi random networks $G(n, p)$ primarily fall in the "low-CNI regime" (Fig. 7), with the value of Θ depending strongly on the average degree $\langle k \rangle = np$ of the network. Simulations suggest that the CNI remains nonnegative, but *can* be zero up until a certain threshold. The CNI reaches a maximum value when $\langle k \rangle \approx 0.33$, but the significance of this value is unclear. Note that most of the interesting features of this graph occur for the Erdős-Rényi graphs without a giant component; when $\langle k \rangle \geqslant 1$, the CNI decreases monotonically as the average degree increases, approaching zero. The behavior of the CNI for a network where many of the nodes with degree zero may be unreliable and is something that needs further study.

B. Barabási-Albert networks

Figure 8(a) shows that Barabási-Albert networks [2] are high-CNI networks, as expected, with a Θ close to the value measured in Fig. 6 for a power-law degree distribution with

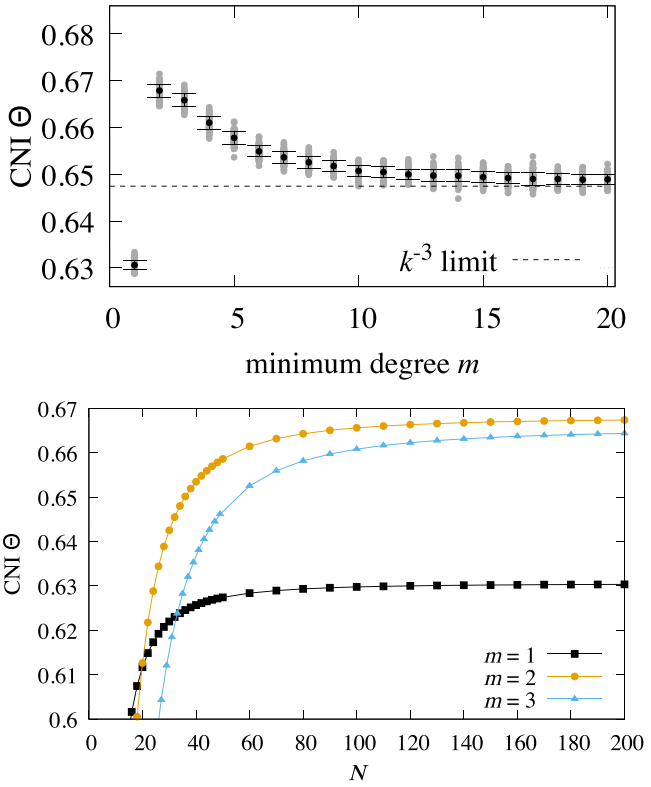

FIG. 8. (a) The CNI for Barabási-Albert networks of 100 000 nodes, as a function of the minimum degree m. The black dots mark the mean value over 100 sampled networks, the error bars show the standard deviation, and the grey dots mark all values. Note the unusual value at $m = 1$. The dashed line shows the CNI of a power-law distribution k^{-3} (0.647, as mentioned in Fig. 6). (b) The partial sum of Eq. (19) for three values of m. Note the clear distinction once again between $m = 1$ and $m = 2$.

$\alpha = 2$. Notice, however, that the CNI depends on the parameter m, which specifies the minimum degree of the network: in particular, the CNI of the $m = 1$ network is noticeably lower than those with higher minimum degrees. This difference seems to contradict the traditional understanding [2] that the infinite-network degree distribution should be $P(k) \propto \frac{1}{k^3}$, independent of the minimum degree m. This might be a finite-size effect, as Barabási-Albert networks are known to converge slowly to their infinite state [22]. To test that theory, we considered the degree distribution $P(k)$ of this network in the thermodynamic limit [23], which is

$$\lim_{N\to\infty} P(k) = \frac{2m(m+1)}{k(k+1)(k+2)}, \; k \geqslant m \tag{18}$$

Combining this with Eq. (16) allows us to write an expression for the CNI which we can numerically estimate:

$$\Theta = \lim_{N\to\infty} \sum_{a=m}^{N} \sum_{b=m}^{N} \sum_{c=m}^{N} \sum_{d=m}^{N} \mathrm{sgn}(\Phi(a, b, c, d)) \times \prod_{s\in\{a,b,c,d\}} \frac{2m(m+1)}{s(s+1)(s+2)}. \tag{19}$$

Figure 8(b) shows the partial sums of Eq. (19) as N approaches infinity. It appears that the CNI for $m = 1$

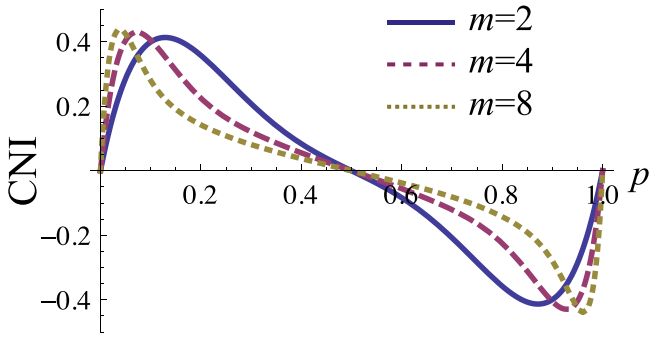

FIG. 9. The CNI of partial periodic lattices with m nearest neighbors, as a function of edge probability p. If at least half of the edges are kept, then the CNI is negative. These correspond to one-dimensional ($m = 2$), two-dimensional ($m = 4$), and four-dimensional ($m = 8$) Cartesian lattices.

approaches a smaller value than for the larger values of m (although it is always possible that it is growing very slowly). Regardless, there is something significantly different about the $m = 1$ case which is being captured by the CNI, likely due to the fact that the CNI is nonasymptotic in nature and depends on more than just the k^{-3} tail. Whether this difference plays a significant role in any applications is worthy of further study.

C. Partial periodic lattices

Another synthetic network which is interesting to us (as we will see in Sec. V A) is what we call a *partial periodic lattice* (PPL), in which each node in a lattice with periodic boundary conditions is connected to each of its m nearest neighbors with probability p. For example, a PPL on a square lattice would have $m = 4$. The CNI of a PPL can be written in closed form as a $4m - 1$ degree polynomial, given by the expression

$$\Theta_{\text{lattice}}(p) = \sum_{i=0}^{m}\sum_{j=0}^{m}\sum_{k=0}^{m}\sum_{l=0}^{m}\operatorname{sgn}(\Phi(i,j,k,l)) \times \prod_{s\in\{i,j,k,l\}}\binom{m}{s}p^s(1-p)^{m-s}. \quad (20)$$

Figure 9 shows this polynomial $\Theta_{\text{lattice}}(p)$ for a few values of m; this looks very similar to the result of the Bernoulli distribution in Fig. 5. We get a negative-CNI result when nodes are connected to more than half of their neighbors.

V. REAL-LIFE NETWORKS

A. Comparison with Broido-Clauset

We now compare our classification scheme for hub-dominant networks to two others proposed in the literature. We begin with the recent paper by Broido and Clauset [4], which considers a set of 927 real-life networks drawn from the ICON database [24]. In that paper they classified each network by how strongly it met the hypothesis that its degree distribution is best fit by a power law. To do so, they used each nonsimple network in their set (i.e., those that are directed, weighted, multipartite, or multiplanar) to generate a collection of unweighted, undirected *simple graphs*, according to criteria

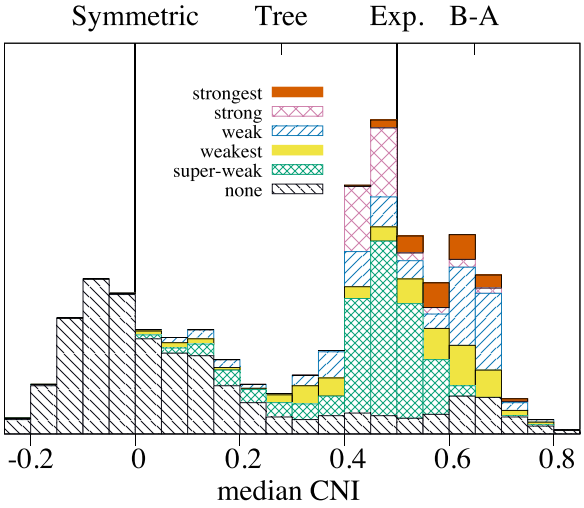

FIG. 10. The distribution of mean CNI for the networks of each strength classification in Ref. [4]. Unlike in that paper, we exclude from the "superweak" category those networks that satisfy the "weakest" condition.

described in Ref. [4]; the *strength* of a network depends on how many of its simple graphs fit the power-law criterion. In the following analysis, we relied on the data provided by Ref. [25].

For each network we define $\bar{\Theta}$ to be the median CNI of that network's collection of simple graphs. Figure 10 shows the distribution of this quantity. The average median CNI for all networks is $\langle\bar{\Theta}\rangle = 0.32$ with a standard deviation of 0.27, but the distribution is bimodal, with peaks around the boundaries of our three classifications (i.e., $\Theta = 0$ and $\Theta = 0.5$). The negative-CNI peak is made up mostly of planar graphs, specifically United States road networks [26] and fungal growth networks [27]: their negative CNI is reminiscent of the partial periodic lattices considered in Section IV C, where nodes are connected to most of their nearest neighbors. Excluding these two outlying groups, the average CNI is $\langle\bar{\Theta}\rangle = 0.49 \pm 0.15$, on the boundary between the high- and low-CNI regimes.

Figure 10 also breaks the distribution down into the strength classifications used in [4], and shows that the two classification schemes are at best only weakly correlated. Most of the strongest fits to the power-law model do have high CNI (though some dip below 0.5, most significantly the protein-protein interaction network in *Mus musculus* [28] with $\Theta = 0.39$). However, 30% of networks in the "weak" category and below are also high-CNI. Overall, 31% of our chosen networks lie in the high-CNI regime; another 24% are close, in the $0.4 \leqslant \Theta < 0.5$ range (suggesting a possible "mid-CNI" regime). Scale-free networks might be rare, as their title suggests, but high-CNI networks are not.

A common way to classify the dominance of hubs in a network is with its tail index α, found by fitting the tail of the degree distribution to a power-law $x^{-\alpha-1}$. Figure 11 shows the CNI of each of our simple graphs versus its tail index as calculated in Ref. [4]. The two values have a moderate negative correlation as one might expect, with a Pearson correlation coefficient of $r = -0.38$. The border between high and low-CNI, according to the fit, occurs at $\alpha = 2.3$, close to the upper range $\alpha = 2$ often cited [4,21] for those networks

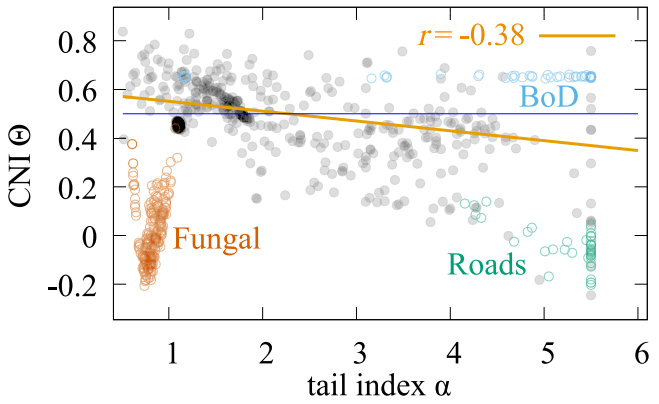

FIG. 11. The tail index of each simple graph versus its CNI, with linear regression line ($\Theta = -0.04\alpha + 0.59$) showing a moderate negative correlation ($r = -0.38$). The line crosses the boundary between high and low-CNI regimes at $\alpha = 2.3$. Three classes of networks are represented with colored open circles: fungal growth networks (red) and US road networks (green) are planar graphs with negative CNI, while the affiliation networks between board directors in Norwegian public limited companies, shown in blue, are further discussed in Fig. 12.

which are "scale-free." However, there are times when the CNI and the tail index differ in surprising ways. For example, the fungal networks and road networks in Fig. 11 are both spatial networks and have similar CNI values even though their tail indices vary greatly. As a more concrete example, consider the set of affiliation networks between board directors on Norwegian public limited companies[29], determined monthly from 2006 through 2009. These networks have a tail index which varies between 1 and 5.5 [see Fig. 12(b)], but their CNI is a fairly constant $\Theta = 0.656 \pm 0.007$ throughout. Do the networks vary significantly or not? If we look at the degree distributions [Fig. 12(a)] from two particular months (May 2006 and August 2006) with very different tail indices ($\alpha = 5.0$ and 1.2, respectively), we see that the two histograms are quite similar, suggesting that the CNI may be a more accurate representation of their heavy-tailed nature.

B. Comparison with Voitalov *et al.*

In response to the assertion in Ref. [4] that scale-free networks are rare, Voitalov *et al.* [6] argues that a finite distribution should be considered "power-law" if it is regularly-varying, and that this more relaxed definition includes the degree sequences of many real-world networks. In their paper, they examine 115 real-world networks from [30], calculating their tail-indices using the Hill [16], Moments [31], and Kernel [32] estimators. They classified these networks into four categories:

(1) *not* power laws: where one tail-index estimate was nonpositive;

(2) *hardly* power laws: where all estimates are positive with at least one estimate $\alpha \geqslant 4$;

(3) other *power laws*: where all estimates are positive, no estimates are $\alpha \geqslant 4$, and at least one estimate is $\alpha \geqslant 2$; and

(4) power laws with *divergent* second-moment: where all estimates are $0 < \alpha < 2$.

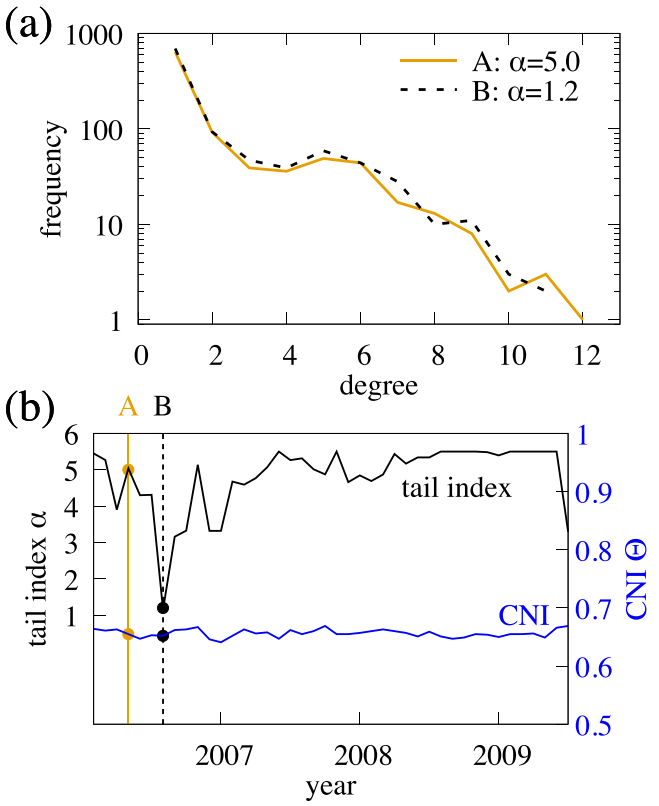

FIG. 12. The top graph shows the degree distribution of the affiliation network between board directors on Norwegian public limited companies [29] in May 2006 (A) and August 2006 (B). While having similar degree distributions, their tail indices α as calculated in [4,25] are very different ($\alpha = 5.0$ and 2.2, respectively). The bottom graph shows how the tail index and CNI of this network varies over time: while the tail index fluctuates widely, the CNI remains relatively stable.

In Fig. 13, we calculate the CNI for the same set of networks as in Ref. [6]. The spread in the CNI for power-law networks is smaller than for non-power-law networks. The mean CNI also increases somewhat as networks become "more power law" (i.e., from category 1 to category 4);

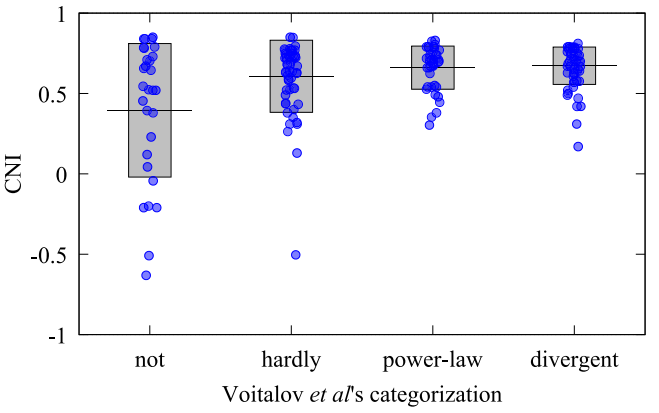

FIG. 13. Each dot represents one of the 115 networks studied in Ref. [6], divided into the four categories mentioned in the text. The horizontal line in each category shows the mean; the grey box, the standard deviation from the mean.

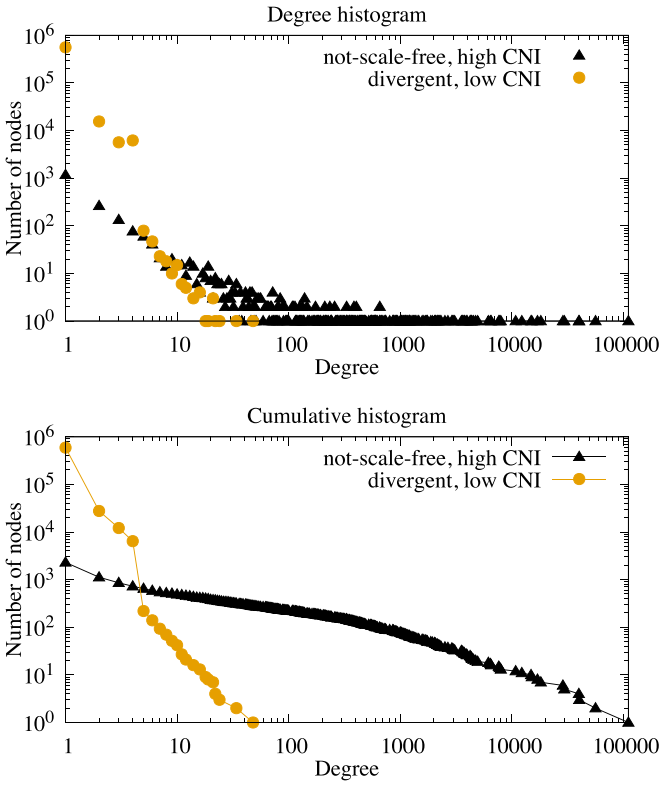

FIG. 14. The histogram (top) and cumulative histograms for the two domains in the CountryDB bipartite network [30,33]. The *countries* domain, marked in triangles, is categorized as "not-scale-free" by Ref. [6] but has a high CNI, while the *entities* domain, in circles, is categorized as "scale-free with a divergent second moment," but has a low CNI.

however, just as in Sec. V A, there are a number of non-power-law networks with high CNI, and a few "divergent" power-law networks with low CNI. An example of both can be found in the bipartite network relating *entities* and their associated *countries* in DBpedia [33]. The degree sequence of the *entities* is classified as scale-free with a divergent second moment by Ref. [6], but has a CNI of 0.17, placing it in the low-CNI category. Conversely, the degree sequence of the *countries* in that network is classified as "not-scale-free" and yet has a high CNI of 0.85. Clearly the CNI and the Voitalov classification are measuring different things. Figure 14 shows the histograms of both sequences, and while the divergent sequence does seem to have a cleaner power-law tail, the high-CNI sequence extends much farther and subjectively seems more hub-dominant than the other.

VI. CONCLUSION

We have introduced the Cooke-Nieboer index as a new and potentially useful method for characterizing hub-dominated networks. The CNI classifies networks into one of three categories: high-CNI which includes the traditional "scale-free" networks and other networks with heavy tails, low-CNI which includes random and regular networks, and negative-CNI which includes planar networks which are mostly connected. While presented here in the context of simple graphs, it is easily generalized to apply to weighted and directed networks. We have shown in Sec. V that our measure is loosely correlated with various other classification schemes, but with some significant differences, due to its nonasymptotic nature. This even occurs with the Barabási-Albert model (Fig. 8), where the CNI makes a distinction between the $m = 1$ and $m = 2$ cases that the tail index would not. We believe the CNI can serve a complementary role in classifying networks as hub-dominated, and encourage its application in the study of epidemics, network fragility, and other fields where the distinction between a power-law network and a hub-dominated network may be important. We also hope that this paper may help network researchers sidestep the controversy over "scale-free networks," when all they care about are hubs.

```python
from random import choices
def cni(degrees, maxerr=0.01):
    S2 = maxerr*maxerr
    Z,D,T = 0,0,0
    while True:
        samp = random.choices(degrees,k=4)
        val = max(samp)+min(samp)-0.5*sum(samp)
        if val > 0:
            D += 1
        elif val < 0:
            D -= 1
        else:
            Z += 1
        T += 1
        if not T%20: #only check every 20 steps
            if T*T - Z*T - D**2 < S2 * T**3:
                return D/T
```

FIG. 15. A more efficient method of estimating the CNI, written in PYTHON.

ACKNOWLEDGMENT

We thank Anna Broido, Aaron Clauset, Phil Chodrow, and Nicole Eikmeier for useful conversations.

APPENDIX

1. An efficient CNI algorithm

The code in Fig. 2 is simple, but computationally inefficient. One can improve the speed somewhat by implementing a running standard error, such as with WELFORD's online algorithm [34]. However, one can do even better by exploiting the fact that the thing we're taking the average of—that is $\mathrm{sgn}(\Phi)$—only takes one of three values. Suppose we take T sets of quadruples from our distribution and calculate $x_i = \mathrm{sgn}\,\Phi_i$ for each one. If we define $D \equiv \sum_i x_i$, then the CNI is $\Theta = D/T$. The variance of this measurement is $\sigma^2 = \frac{1}{T}\sum_i x_i^2 - \langle x_i\rangle^2$. Because x_i^2 is either zero or one, $\sum_i x_i^2 = T - Z$ where Z is the number of times that $\Phi_i = 0$. Thus the variance can be written

$$\sigma^2 = \frac{T-Z}{T} - \left(\frac{D}{T}\right)^2 = 1 - \frac{ZT + D^2}{T^2} \tag{A1}$$

and thus the squared standard error is

$$\sigma_{\bar{x}}^2 = \frac{1}{T}\sigma^2 = \frac{1}{T} - \frac{ZT + D^2}{T^3}. \tag{A2}$$

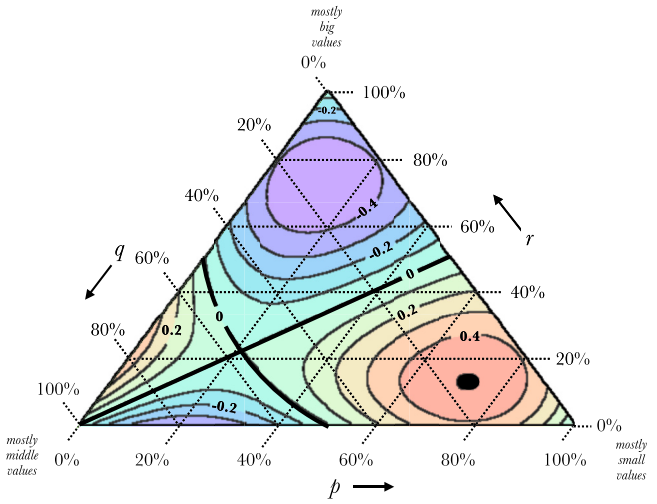

FIG. 16. The CNI of the "trinoulli" distribution defined in Eq. (A3), with $j = 0$ (that is, $c = 2b - a$). The CNI is the most positive when small values predominate (lower-right corner, $p \gg q, r$), and most negative when large values predominate (top, $r \gg p, q$). The black dot in the lower-right corner are those values where $\Theta > 0.5$. Figures for $j = \pm 1$ are similar, with slightly higher values for CNI overall when $c > 2b - a$, and slightly lower values when $c < 2b - a$.

This confirms the result seen in Fig. 4 that $\frac{1}{\sqrt{T}}$ is an upper-bound and a good approximation for $\sigma_{\bar{x}}$, so long as Z and D are both much smaller than T.

The code in Fig. 15 uses this insight to determine the standard error, and calculates the CNI almost 3 times faster than code using the WELFORD algorithm, and 75 times faster than the code in Fig. 2.

2. Star networks and trinoulli distributions

Because the obesity index was originally designed for continuous distributions, it is unsurprising that it may have difficulty with distributions with few unique elements. For example, a star graph, consisting of one hub and N nodes, is a classical example of a hub-dominated network. However, as its degree sequences follows a Bernoulli distribution with $a = 1$ and $b = N$, it cannot have a CNI larger than the maximum value for a Bernoulli distribution, which according to Sec. III A is $\Theta = 0.385$, making it a low-CNI network. Of course, as the histogram of this network consists of two points, defining its tail-index is also problematic. Note that a "tri-noulli distribution," defined as

$$X = \begin{cases} a & \text{with probability } p \\ b > a & \text{with probability } q \\ c > b & \text{with probability } r = 1 - p - q \end{cases} \tag{A3}$$

has

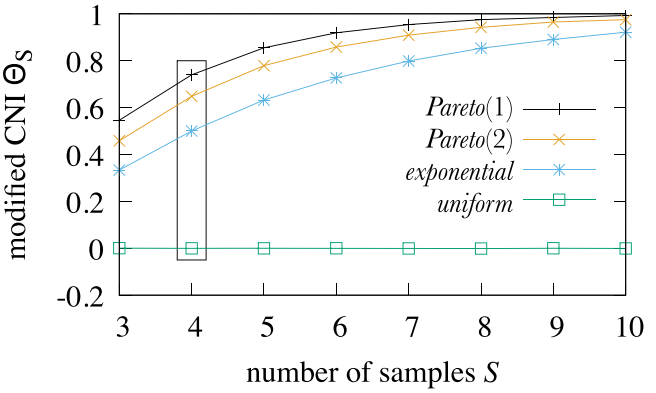

FIG. 17. The generalized CNI using an S–tuple, for four different distributions: the Pareto distributions with $\alpha = 1$ and 2 (x^{-2} and x^{-3}, respectively), the exponential distribution $e^{-\lambda x}$, and a uniform distribution of numbers between 0 and 1. The blue exponential curve would mark the boundary between high and low CNI.

$$\Theta = 4[p^3(q + r) + q^3(r - p) - (p + q)r^3 + 3pqr(p - r + jq)], \tag{A4}$$

where $j = \text{sgn}(c - 2b + a)$. This *can* (barely) reach the high-CNI regime, as is shown in Fig 16.

3. Using other-sized tuples

Equation (8) specifies the use of a quadruple when calculating Φ, but Eq. (10) could be interpreted to refer to any S-tuple, and the resulting modified CNI would be different. Figure 17 shows this modified CNI Θ_S for several basic continuous distributions. The value for a uniform distribution remains zero throughout, but for others, Θ_S increases monotonically as the size S of the tuple increases, compressing the "high-CNI" regime and expanding the "low-CNI" regime. We use $S = 4$ in this paper not only to maintain continuity with [14], but because it gives the exponential distribution a CNI of 0.5, and so divides the positive range of values evenly between the high and low regimes. Note that, while the ordering of the example distributions in Fig. 17 does not change with S, this is not true in general. For example, the modified CNI of the Bernoulli distribution [Eq. (13)] for an S-tuple is

$$\Theta_S(p) = \sum_{z=1}^{\lceil S/2-1 \rceil} \binom{S}{z} [p^{S-z}(1-p)^z - p^z(1-p)^{S-z}] \tag{A5}$$

and one can show that, for example, $\Theta_4(0.77) < \Theta_4(0.79)$ but $\Theta_7(0.77) > \Theta_7(0.79)$. This suggests it may be possible that different values of S may result in different classifications for certain networks, a possibility that may be worth further study.

[1] P. Erdős and A. Rényi, On Random Graphs I, Publications Mathematicae Debrecen **6**, 290 (1959).

[2] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).

[3] M. E. J. Newman, *Networks: An Introduction*, 1st ed. (Oxford University Press, Oxford, New York, 2010).
[4] A. D. Broido and A. Clauset, Scale-free networks are rare, Nat. Commun. **10**, 1017 (2019).
[5] A.-L. Barabási, *Network Science* (Cambridge University Press, 2016).
[6] I. Voitalov, P. van der Hoorn, R. van der Hofstad, and D. Krioukov, Scale-free Networks Well Done, Phys. Rev. Research **1**, 033034 (2019).
[7] A. Clauset, C. R. Shalizi, and M. E. Newman, Power-law distributions in empirical data, SIAM Rev. **51**, 661 (2009).
[8] C. Song, S. Havlin, and H. Makse, Self-similarity of complex networks, Nature (London) **433**, 392 (2005).
[9] M.Á.Serrano, D. Krioukov, and M. Boguñá, Self-Similarity of Complex Networks and Hidden Metric Spaces, Phys. Rev. Lett. **100**, 078701 (2008).
[10] A.-L. Barabási, Love is all you need, https://www.barabasilab.com/post/love-is-all-you-need (2018).
[11] P. Holme, Rare and everywhere: Perspectives on scale-free networks, Nat. Commun. **10**, 1016 (2019).
[12] S. Foss, D. Korshunov, and S. Zachary, *An Introduction to Heavy-Tailed and Subexponential Distributions*, 2nd ed., edited by T. V. Mikosch, S. I. Resnick, and S. M. Robinson, Springer Series in Operations Research and Financial Engineering (Springer, 2013).
[13] R. Pastor-Satorras and A. Vespignani, Epidemic Spreading in Scale-Free Networks, Phys. Rev. Lett. **86**, 3200 (2001).
[14] R. M. Cooke, D. Nieboer, and J. Misiewicz, *Fat-Tailed Distributions: Data, Diagnostics, and Dependence*, Mathematical Models and Methods in Reliability Set No. 1 (Wiley-ISTE, 2014).
[15] S. Kotz and S. Nadarajah, *Extreme Value Distributions: Theory and Applications* (Imperial College Press, London, 2000).
[16] B. M. Hill, A simple general approach to inference about the tail of a distribution, Ann. Stat. **3**, 1163 (1975).
[17] J. Pickands III, Statistical inference using extreme order statistics, Ann. Stat. **3**, 119 (1975).
[18] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, Resilience of the Internet to Random Breakdowns, Phys. Rev. Lett. **85**, 4626 (2000).
[19] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, Breakdown of the Internet under Intentional Attack, Phys. Rev. Lett. **86**, 3682 (2001).
[20] C. M. Goldie and C. Klüppelberg, Subexponential distributions, in *A Practical Guide to Heavy Tails: Statistical Techniques and Applications* (1998), pp. 435–459.
[21] S. Dorogovtsev and J. Mendes, Evolution of networks, Adv. Phys. **51**, 1079 (2002).
[22] B. Waclaw and I. M. Sokolov, Finite size effects in barabási-albert growing networks, Phys. Rev. E **75**, 056114 (2007).
[23] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, The degree sequence of a scale-free random graph process, Random Struct. Alg. **18**, 279 (2001).
[24] A. Clauset, E. Tucker, and M. Sainz, The Colorado Index of Complex Networks, https://icon.colorado.edu (2016).
[25] A. D. Broido and A. Clauset, Scale-free network analysis, https://github.com/adbroido/SFAnalysis (2019).
[26] D. Schultes, United States Road Networks (TIGER/Line), http://www.dis.uniroma1.it/challenge9/data/tiger/ (2005).
[27] S. Lee, M. Fricker, and M. Porter, Mesoscale analyses of fungal networks, J. Complex Networks **5**, 145 (2017).
[28] J. Das and H. Yu, Hint: High-quality protein interactomes and their applications in understanding disease, BMC Systems Biology **6**, 92 (2012).
[29] C. Seierstad and T. Opsahl, For the few not the many!, Scandanavian Journal of Management **27**, 44 (2011).
[30] J. Kunegis, KONECT–The Koblenz Network Collection, in *Proceedings of the International Conference on World Wide Web Companion* (Association for Computing Machinery, 2013), pp. 1343–1350.
[31] A. Dekkers, J. Einmahl, and L. De Haan, A moment estimator for the index of an extreme-value distribution, Ann. Stat. **17**, 1833 (1989).
[32] P. Groeneboom, H. Lopuhaä, and P. De Wolf, Kernel-type estimators for the extreme value index, Ann. Stat. **31**, 1956 (2003).
[33] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, Dbpedia: A nucleus for a web of open data, in *Proceeding of International Semantic Web Conference* (Springer-Verlag, 2008), pp. 722–735.
[34] B. Welford, Note on a method for calculating corrected sums of squares and products, Technometrics **4**, 419 (1962).