A nonasymptotic measure for characterizing heavy-tailed networks

Scott A. Hill*
Adrian College, Adrian MI

Heavy-tailed networks are often characterized in the literature by their degree distribution's similarity to a power law. However, many heavy-tailed networks in real life do not have power-law degree distributions, and in many applications the scale-free nature of the network is irrelevant so long as the network possesses hubs. In this presentation we introduce the Cooke-Nieboer index (CNI), a non-asymptotic measure of the heavy-tailedness of a network's degree distribution which does not presume a power-law form. We apply the measure to several synthetic and real-life networks, compare it to the standard tail index method, and discuss its ability to distinguish between heavy-tailed, random, and planar networks.

I. MOTIVATION

Modern network science was born with the discovery that the relationships in many real-life systems could not be described via the random graph theory of Erdős and Rényi[1]. In particular, many real-life networks have *hubs*, or nodes with degrees much larger than a random network of the same size and average degree would possess. Because the degree distributions of such networks extend much farther than the typical Poisson distribution of Erdős-Rényi graphs, these networks are called *heavy-tailed*. The most well-known model with hubs is the Barabasi-Albert model[2]. The Barabasi-Albert network has a "scale-free" or power-law degree distribution ($P(x) \propto x^{-\alpha}$), which corresponded with several large networks which were studied by the authors. It thus became common in the network science community to describe heavy-tailed networks as "scale-free networks", so that the two terms have become synonymsz, even in cases when a power-law is only a rough approximation for a network's degree distribution.

In recent years, this conflation has come under greater scrutiny[3]. Many real-world networks have been shown to have distributions which are better described as log-normal or Weibull distributions[3]. Even a finite Barabasi-Albert network is not strictly "scale-free", as it is bounded by the size of the network. In some sense this is a semantic controversy, as the description "scale-free" is understood to be approximate. On the other hand, there may be some advantages in distinguishing between heavy-tailed and scale-free networks. For example, the proof[4] that scale-free networks with tail index $\alpha < 3$ have no epidemic threshold relies on the fact that such power-law distributions have infinite variance. However, epidemics might spread quite differently on a network with a log-normal degree distribution, which has finite variance[5].

In many lines of research, it is the presence of hubs, rather than the actual shape of the degree distribution, which is significant. In such cases, it is common to define the *tail index* α of a network by calculating the average slope of a portion of its degree distribution on a log-log plot, as if it were a power-law. There are problems with this approach, however. Selecting the portion of the slope to fit can be nontrivial. A naive linear fit on a log-log plot can introduce significant systematic errors[6]. Furthermore, there is the philosophical difficulty of defining a measure for heavy-tailed networks based on a distributional form which is only possessed by a minority of real-life networks[3].

In this paper we present a new quantity, called the *Cooke-Nieboer index*, which measures the heavy-tailedness of a network without presuming it to be scale-free. We calculate its value for several theoretical distributions and synthetic networks to understand its properties, and then apply it to a number of real-world networks from the ICON database[7], discussing its relationship with the tail index α and its "strength" as discussed in [3]. We will conclude with a peculiarity of the measure when it comes to star graphs.

II. THE OBESITY INDEX

In the probability literature[8], a distribution $F(x)$ is said to be *heavy-tailed* if

$$\int_{-\infty}^{\infty} e^{\lambda x} F(x)\, dx = \infty \quad \text{for all } \lambda > 0. \tag{1}$$

Most heavy-tailed distributions of interest are *subexponential*: if $X_1, \ldots, X_n$ are independent and identically distributed (iid) random variables chosen from a subexponential distribution, then[9]

$$\lim_{x\to\infty} \frac{P(X_1 + \cdots + X_n > x)}{P(\max(X_1, \ldots, X_n) > x)} = 1, \text{ for all } n \geq 1 \tag{2}$$

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*[8]. (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

* shill@adrian.edu

that year.) To measure the "subexponentiality" of a distribution X, Cooke and Nieboer[10] defined its *obesity index* as follows: select four random values from the distribution and label them in ascending order, so that $X_1 \le X_2 \le X_3 \le X_4$. Then

$$\mathrm{Ob}(X) = P(X_4 + X_1 > X_2 + X_3) \tag{3}$$

That is, if the distribution is subexponential, then X_4 will probably be larger than the other three variables combined, and so $X_1 + X_4$ must certainly be greater than $X_2 + X_3$. On the other hand, for a symmetric distribution the values $X_4 + X_1$ and $X_2 + X_3$ are equally likely to be larger, and so the obesity index of a symmetric distribution is one-half.[10]

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

$$\mathrm{Ob}(aX + b) = \mathrm{Ob}(X),\ a \in \mathbb{R}^+, b \in \mathbb{R} \tag{4}$$

The obesity index does not depend on asymptotic behavior alone, as does the tail index, but takes the entire distribution into account.

III. THE COOKE-NIEBOER INDEX

Starting with the obesity index, we define the *Cooke-Nieboer index* (CNI) of a distribution, which we represent by the symbol Θ. The CNI differs from the obesity index in three ways: (i) The CNI ranges from -1 to 1, so that the CNI of symmetric distributions is $\Theta = 0$; (ii) our measure is defined in terms of discrete distributions, by accounting for the finite possibility that $X_1 + X_4 = X_2 + X_3$; and (iii) we avoid the term "obesity index", which may cause confusion in network science studies of health issues.

Definition: Let $X_1, \ldots, X_4$ be four iid random values chosen from a particular distribution X, such as the degree distribution of a network. We define

$$\Theta(X) \equiv E\left\{\mathrm{sgn}\left(2(\max X_i + \min X_i) - \sum_i X_i\right)\right\}, \tag{5}$$

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

$$\mathrm{sgn}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}. \tag{6}$$

For later convenience, we also define

$$\Phi(X) = 2(\max X_i + \min X_i) - \sum_i X_i \tag{7}$$

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

For a finite distribution with N data points, the exact CNI can be calculated in $O(N^4)$ time, by considering every set of four points. (Note that the requirement that the values be independent means that the selection is made with replacement.) Typically, however, we will use a Monte Carlo simulation such as the one in Fig. 1, which can calculate the CNI to whatever accuracy we require.

```
import numpy as np
import random
def cni(degrees,maxerr=0.0001):
   vals=[]
   while True:
      #Choose four samples from degrees
      four=random.choices(degrees,k=4)
      val=max(four)+min(four)-0.5*sum(four)
      vals+=[np.sign(val)]
      sterr=np.std(vals)/np.sqrt(len(vals))
      if(len(vals)>10 and sterr<maxerr):
         return np.mean(vals)
```

FIG. 1. Sample Python code for calculating the CNI, given a list `degrees` of degrees of the network. In practice this algorithm can be sped up with a running standard error algorithm such as the Welford algorithm[11], not shown. Experience suggests that the program needs to draw $2.7/(SE)^2$ sets of samples to calculate an CNI with standard error SE, regardless of the size of the network.

IV. DISTRIBUTIONS

Because the CNI of a distribution is simply related to its obesity index

$$\Theta(X) = 2\,\mathrm{Ob}(X) - 1 \tag{8}$$

we can find the CNI of some continuous distributions that were studied in [10]. Cooke and Nieboer showed that the exponential distribution $P(x) = \lambda e^{-\lambda x}$ has an CNI of 0.5, while the power-law distribution $P(x) \propto x^{-\alpha}$ has an CNI between 0.5 (as $\alpha \to \infty$ and 1 (as $\alpha \to 1$) (Fig. 2).

We will divide distributions into three categories:

- **High-CNI** distributions, with $\Theta > 0.5$. These are the *subexponential* distributions, and include the power-law, Wiebull ($x^{k-1}e^{-x^k}$ with $k < 1$), and lognormal ($e^{-(\ln x - \mu)^2/2\sigma^2}$ with $\sigma \gtrsim 1$) distributions.
- **Low-CNI** distributions, with $0 \le \Theta \le 0.5$. These include the symmetric distributions, as well as the Poisson distribution (as seen in Fig. 4).
- **Negative-CNI** distributions, with $\Theta < 0$.

To see an example of this third category, we consider the *bimodal distribution*

$$X = \begin{cases} a & \text{with probability } p \\ b > a & \text{with probability } 1 - p \end{cases}. \tag{9}$$

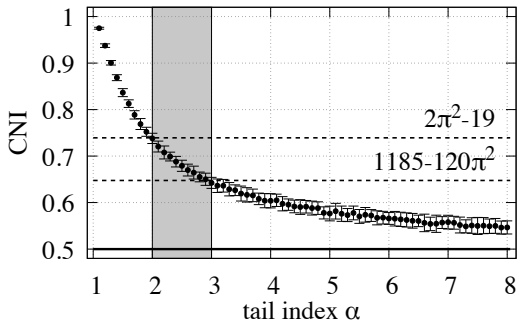

FIG. 2. The CNI of a power-law distribution $x^{-\alpha}$ as a function of its tail index α, calculated via numerical simulation. Each dot is the average CNI over 50 different sets of 1000 samples from the distribution, with error bars indicating the standard deviation. The grey area highlights the region where most "scale-free" networks are found, between $\alpha = 2$ and $\alpha = 3$. Ref. [10] calculates the CNI at these values as $2\pi^2 - 19$ and $1185 - 120\pi^2$, respectively.

If we choose four numbers from this distribution, and $0 \le s \le 4$ of them are a, it is simple to show that Φ (Eq. 7) is equal to zero if s is even, $\Phi < 0$ if $s = 1$, and $\Phi > 0$ if $s = 3$. Thus we can calculate the CNI of this distribution precisely:

$$\begin{aligned}\Theta(p) &= \sum_{s=0}^{4} \binom{4}{s} p^s (1-p)^{4-s} \operatorname{sgn}(\Phi) \\ &= 4p^3(1-p) - 4p(1-p)^3 \\ &= 4p(1-p)(2p-1).\end{aligned} \tag{10}$$

Note that the result does not depend on the values a and b. Fig. 3 shows a graph of this polynomial. The distribution is symmetric when $p = 0$, $p = 0.5$, and $p = 1$, and so the CNI is zero. When the smaller values are predominant, as in typical degree distributions, the CNI is positive, with a maximum value of $\frac{2\sqrt{3}}{9} \approx 0.385$ (and thus never in the "high regime") at $p = \frac{1}{2} + \frac{\sqrt{3}}{6} \approx 0.79$. (So about 1 of every 5 nodes should be hubs for maximal CNI). When the larger values are predominant, however, the CNI can be negative.

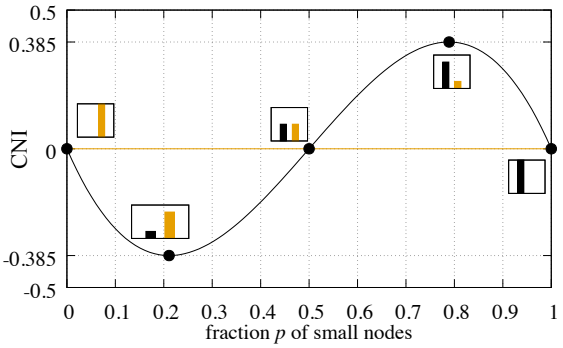

FIG. 3. The CNI of the bimodal distribution (Eq. 11) as a function of p. The small boxes show the relative proportions of the two values ($X = 0$ in black, $X = 1$ in orange). The polynomial reaches extreme values of $\pm\frac{2\sqrt{3}}{9}$ at $p = \frac{1}{2} \pm \frac{\sqrt{3}}{6}$.

Similarly for a *trimodal distribution*,

$$X = \begin{cases} a & \text{with probability } p \\ b > a & \text{with probability } q \\ c > b & \text{with probability } r = 1 - p - q \end{cases}. \tag{11}$$

we can calculate the CNI to be

$$\Theta = 4[p^3(q+r) + q^3(r-p) - (p+q)r^3 + 3pqr(p-r+jq)]$$

where $j = \operatorname{sgn}(c - 2b + a)$. Trimodal distributions *can* reach the high-CNI regime; for example, $\Theta\left(p = \frac{2}{3}, q = r = \frac{1}{6}\right) = \frac{14}{27} = 0.52$.

V. SYNTHETIC NETWORKS

We define $\Theta(G)$ for an undirected, unweighted network G to be the CNI of its degree distribution; that is, $\Theta(G) = E\{\operatorname{sgn}(\Phi)\}$ where

$$\Phi = 2(\max k_{n_i} + \min k_{n_i}) - \sum_i k_{n_i},\ n_1, \ldots, n_4 \in G \tag{12}$$

and k_{n_i} is the degree of node n_i in G. Regular networks such as complete graphs K_n and cycle graphs C_n, have $\Theta = 0$, as do networks with symmetric degree distributions. A star graph S_n, a tree with one internal node and n leaves has a bimodal distribution with $p = \frac{n}{n+1}$, and so (Eq. 11) $\Theta(S_n) = 4\frac{n(n-1)}{(n+1)^3}$, with a maximum value of 0.384 for a star S_4 with four leaves. We will discuss this rather peculiar result in Section VII.

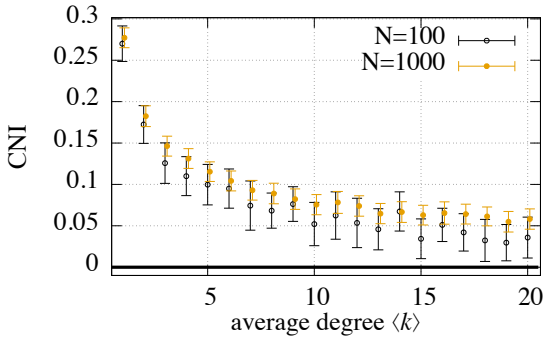

FIG. 4. The mean CNI of Erdős-Rényi networks with 100 and 1000 nodes, as a function of the average degree $\langle k \rangle$, averaged over 50 networks for each dot. The error bars indicate the standard deviation, and the $N = 1000$ dots are slightly offset for clarity.

The Erdős-Rényi random networks[1], where every pair of nodes are connected with probability $p = \langle k \rangle / N$,

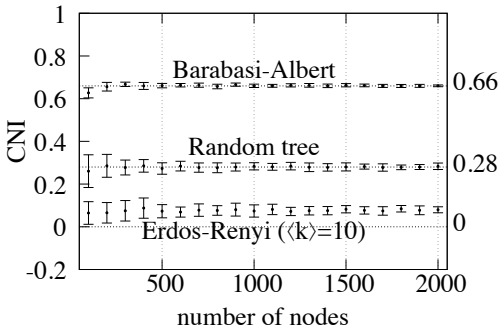

FIG. 5. The CNI for Erdős-Rényi networks, random trees, and Barabasi-Albert networks having different numbers of nodes. The Barabasi-Albert and Erdős-Rényi graphs both have average degree $\langle k\rangle = 10$. The error bars show the variation (standard deviation) in the CNI when generated over 20 samples of each network.

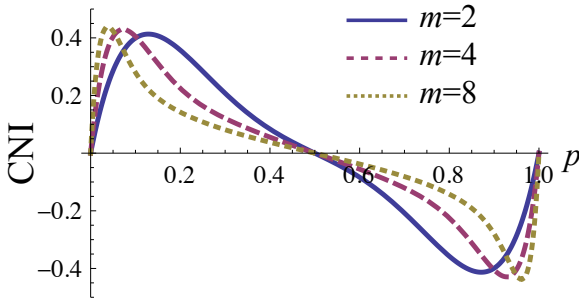

FIG. 6. The CNI of partial periodic lattices with m nearest neighbors, as a function of edge probability p. If at least half of the edges are kept, then the CNI is negative.

have a low but positive CNI which depends on the average degree of the network (Fig. 4). Barabasi-Albert networks[2], on the other hand, are high-CNI networks with $\Theta = 0.66$, consistent with the power-law degree distribution x^{-3} seen in Fig. 2. The CNI of both networks is independent of the number of nodes N for sufficiently large N (Fig. 5).

Another interesting example is a *partial periodic lattice*, in which each node in a lattice is connected to each of its m nearest neighbors with probability p. If $p = 1$, each node has the same degree m, and the CNI is zero. In general,

$$\Theta_{\text{lattice}}(p) = \sum_{i=0}^{m}\sum_{j=0}^{m}\sum_{k=0}^{m}\sum_{l=0}^{m}\text{sgn}(\Phi(i,j,k,l)) \times \prod_{s\in\{i,j,k,l\}}\binom{m}{s}p^s(1-p)^{m-s}. \quad (13)$$

Figure 6 shows $\Theta_{\text{lattice}}(p)$ for a few values of m. Generally, this is a $(4m-1)$–degree polynomial which is negative when $p > 0.5$, and approximately $\Theta \approx -4m(1-p)$ when $p \approx 1$. This critical value $p = 0.5$ is also the bond percolation threshold of the square lattice[12], so square partial periodic lattices with a giant cluster are examples of negative-CNI networks (although this connection with percolation is, as far as we know, coincidental).

VI. REAL-LIFE NETWORKS

We now apply our measure to a set of 927 real-life networks referred to and analyzed in [3], drawn from the ICON database[7]. Each non-simple network (i.e. those that are directed, weighted, multipartite, or multiplanar) is used to generate a collection of unweighted, undirected *simple graphs*, according to criteria described in [3]. We define $\bar{\Theta}$ of a network to be the median CNI of the network's collection of graphs.

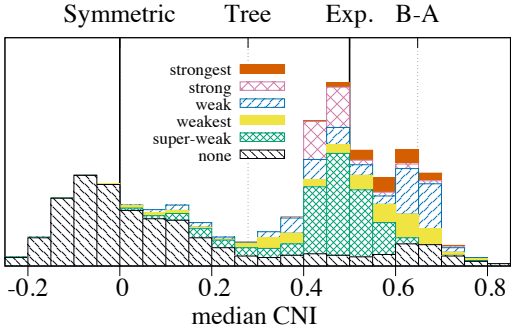

FIG. 7. The distribution of mean CNI for the networks of each strength classification. Unlike in [3], we exclude from the "super-weak" category those networks that satisfy the "weakest" condition.

Figure 7 shows the distribution of the networks' median CNI, $\bar{\Theta}$. The average median CNI over all networks is $\langle\bar{\Theta}\rangle = 0.32\pm0.27$, but the distribution is bimodal, with one peak around $\Theta = 0.5$ and one just below $\Theta = 0$. The negative-CNI peak is made up mostly of planar graphs, specifically United States road networks[13] and fungal growth networks[14]. Their negative CNI is reminiscent of the partial periodic lattices considered in the previous section, and it may be that $\Theta < 0$ is a signature of real-world planar networks. Excluding these two outlying groups, the average CNI is $\langle\bar{\Theta}\rangle = 0.49 \pm 0.15$. Fig. 7 also breaks the distribution down into the strength classifications used in [3], according to how strong a fit each was to a power-law degree distribution. Most of the networks which most strongly fit the power-law model have high CNI, with a few low-CNI exceptions. However, 30% of networks in the "weak" category and below are also high-CNI. Overall, 31% of these networks lie in the high-CNI regime, qualifying as "subexponential"; another 24% are close, in the $0.4 \le \Theta < 0.5$ range (which suggests a new "mid-CNI" regime). While scale-free networks may be rare, high-CNI networks are not.

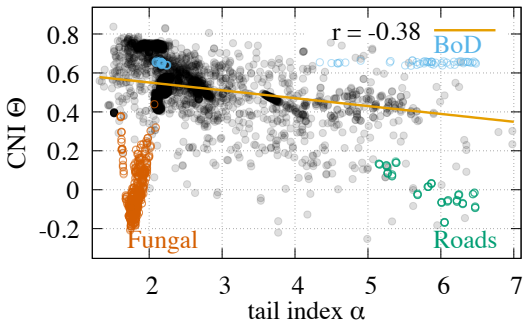

FIG. 8. The tail index of each simple graph versus its CNI, with linear regression line showing a moderate negative correlation ($r = -0.38$). Three classes of networks are represented with colored open circles: fungal growth networks (red) and US road networks (green) are planar graphs with negative CNI, while the affiliation networks between board directors in Norwegian public limited companies, shown in blue, are further discussed in Fig. 9.

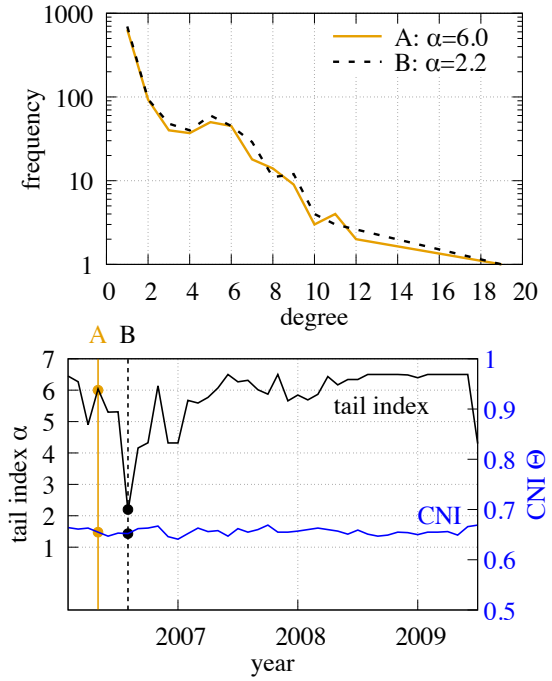

FIG. 9. The top graph shows the degree distribution of the network representing the affiliation network between board directors on Norwegian public limited companies[15] in May 2006 (A) and August 2006 (B). While having similar degree distributions, their tail indices α are very different ($\alpha = 6.0$ and $\alpha = 2.2$, respectively). The bottom graph shows how the tail index and CNI of this network varies over time: while the tail index fluctuates widely, while the CNI remains relatively stable.

Another way to classify the heaviness of a network's tail is with its tail index α, found by fitting the tail of the degree distribution to a power-law $x^{-\alpha}$, whether or not a power-law fit is appropriate[3, 6]. Networks traditionally considered to be "scale-free" have a tail index $2 < \alpha < 3$. Figure 8 shows the CNI of each of our simple graphs versus its tail index: the two values have a moderate negative correlation as one might expect, with a Pearson correlation coefficient of $r = -0.38$.

However, there are times when the two quantities differ in surprising ways. Consider the set of affiliation networks between board directors on Norwegian public limited companies[15], determined monthly from 2006 through 2009. These networks have a tail index which varies between 2 and 6.5 (see the inset to Fig. 9), 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. 9) from two particular months (May 2006 and August 2006) with very different tail indices ($\alpha = 6.0$ and $\alpha = 2.2$, respectively), we see that the two histograms are much more similar than the tail index would suggest.

VII. THE STAR-GRAPH ANOMALY

When switching from continuous to discrete degree distributions, as we've done here, it is not unusual for problems to arise. For example, consider the star graph S_n, which consists of one hub with degree n attached to n leaves of degree 1. (FIG. 10) While it would be a stretch to call this a "scale-free" network due to its bimodal degree distribution, one cannot deny that it possesses a hub; and in fact, if you naively fit its degree distribution to a power-law you end up with a tail index of $\alpha = 1$. However, because this is a bimodal degree distribution, we can calculate its CNI $\Theta(S_n) = \frac{4n(n-1)}{(n+1)^3}$ using Eq. 10, and we already know from Fig. 3 that $\Theta < 0.385$ for this network. Not only are the star graphs not "subexponential" according to our measure, but their CNI approaches zero as n gets large (Fig. 10).

The problem lies in the fact that Eq. 10 does not depend on the actual degrees of the two types of nodes, only their numbers. Thus, when a star graph gains a leaf, the single hub increases in degree, but this increase has no effect on the CNI of the graph; meanwhile the number of leaves increase, making the hub more of an anomaly, and so the CNI decreases once the ratio of leaves to hubs exceeds the optimal four. This anomaly occurs when there is a large gap in the distribution, with a range of degrees having no representation in the network. A revised version of this measure must take this gap into account somehow. Note that a cycle of N S_n star graphs (Fig. 10b) has the same limitation; in fact, removing a single leaf from this chain will actually increase the CNI, even though one of the hubs becomes smaller as a result.

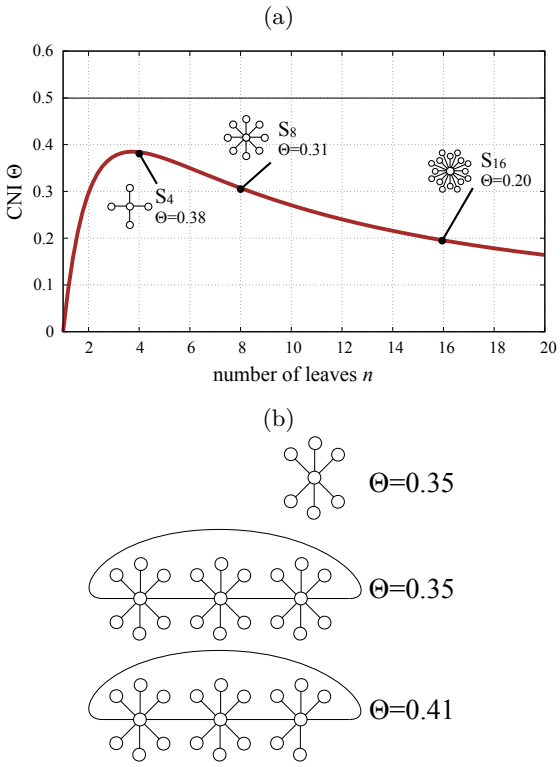

FIG. 10. (a) The CNI of a star graph S_n as a function of its number of leaves. A star graph with four leaves has the largest CNI, but none of them are in the high-CNI regime $\Theta > 0.5$. (b) This chain of $N = 3$ S_6 star graphs has the same CNI as a single S_6 star graph; i.e. $\Theta = 0.35$. Removing one leaf from the rightmost star, however, increases the CNI to $\Theta = 0.41$.

VIII. CONCLUSION

We have introduced the Cooke-Nieboer index as a potentially useful method for classifying heavy-tailed networks. It would be interesting to see if high-CNI networks share the same characteristics with power-law networks, such as a small diameter, robustness to failure, and a vanishing epidemic threshold.

In the future we would like to look at variations of the CNI measure. It is trivial, for instance, to replace node degrees with node strengths in 1, and apply the CNI to weighted networks; whether this weighted CNI has physical significance is to be determined. The choice of four samples to calculate the CNI is somewhat arbitrary, and Eq. 5 as written can allow for a different number of samples. A revised version might also be found which eliminates the star-graph anomaly, or perhaps there is some justification as to why star graphs should not be considered "heavy-tailed" networks at all. The naive algorithm to calculate the precise CNI of a network with N nodes is $O(N^4)$, but there may be a faster algorithm which would eliminate the need for a Monte Carlo approach.

Primarily, however, we would like to see this model applied in situations where the heavy-tailed or scale-free property of networks have been shown to be important: in epidemiology[4], network fragility[16], and so forth. We would like to know whether this measure correlates with other features associated with heavy-tailed networks, and whether the classification of "high-CNI" might end up being more useful than whether a network is "scale-free" or not.

We thank Anne Broido and Aaron Clauset for making their data available in a convenient format at https://github.com/adbroido/SFAnalysis; we relied heavily on their data in Section VI.

[1] P. Erdős and A. Rényi, Publicationes Mathematicae **6**, 290 (1959).
[2] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[3] A. D. Broido and A. Clauset, Nature Communications **10**, 1017 (2019).
[4] R. Pastor-Satorras and A. Vespignani, Physical Review Letters **86**, 3200 (2001).
[5] J. K. Blitzstein and J. Hwang, *Introduction to Probability* (CRC Press, 2019).
[6] A. Clauset, C. R. Shalizi, and M. E. Newman, SIAM Review **51**, 661 (2009).
[7] A. Clauset, E. Tucker, and M. Sainz, "The Colorado Index of Complex Networks," `https://icon.colorado.edu`.
[8] 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).
[9] C. M. Goldie and C. Klüppelberg, in *A practical guide to heavy tails: statistical techniques and applications* (1998) pp. 435–459.
[10] R. Cooke, D. Nieboer, and J. Misiewicz, *Fat-Tailed Distributions: Data, Diagnostics, and Dependence*, Discussion Paper RFF DP 11-19-REV (Resources for the Future, 2011).
[11] B. Welford, Technometrics **4**, 419 (1962).
[12] H. Kesten, Comm. Math. Phys. **74**, 41 (1980).
[13] D. Schultes, "United States Road Networks (TIGER/Line)," `http://www.dis.uniroma1.it/challenge9/data/tiger/`, October 2005.
[14] S. Lee, M. Fricker, and M. Porter, Journal of Complex Networks **5**, 145 (2017).
[15] C. Seierstad and T. Opsahl, Scandanavian Journal of Management **27**, 44 (2011).
[16] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin, Physical Review Letters **85**, 4626 (2000).