Thresholding normally distributed data creates complex networks

George T. Cantwell,[1,*] Yanchen Liu,[2] Benjamin F. Maier,[3,4] Alice C. Schwarze,[5] Carlos A. Serván,[6] Jordan Snyder,[7,8] and Guillaume St-Onge[9,10]

[1] *Department of Physics, University of Michigan, Ann Arbor, Michigan 48109, USA*
[2] *Center for Complex Network Research, Northeastern University, Boston, Massachusetts 02115, USA*
[3] *Robert Koch Institute, Nordufer 20, D-13353 Berlin, Germany*
[4] *Department of Physics, Humboldt-University of Berlin, Newtonstraße 15, D-12489 Berlin, Germany*
[5] *Wolfson Centre for Mathematical Biology, Mathematical Institute, University of Oxford, Oxford OX2 6GG, United Kingdom*
[6] *Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA*
[7] *Department of Mathematics, University of California, Davis, California 95616, USA*
[8] *Complexity Sciences Center, University of California, Davis, California 95616, USA*
[9] *Département de Physique, de Génie Physique, et d'Optique, Université Laval, Québec (Québec), Canada G1V 0A6*
[10] *Centre Interdisciplinaire de Modélisation Mathématique de l'Université Laval, Québec (Québec), Canada G1V 0A6*

(Received 6 March 2019; revised manuscript received 6 May 2020; accepted 7 May 2020; published 1 June 2020)

Network data sets are often constructed by some kind of thresholding procedure. The resulting networks frequently possess properties such as heavy-tailed degree distributions, clustering, large connected components, and short average shortest path lengths. These properties are considered typical of complex networks and appear in many contexts, prompting consideration of their universality. Here we introduce a simple model for correlated relational data and study the network ensemble obtained by thresholding it. We find that some, but not all, of the properties associated with complex networks can be seen after thresholding the correlated data, even though the underlying data are not "complex." In particular, we observe heavy-tailed degree distributions, a large numbers of triangles, and short path lengths, while we do not observe nonvanishing clustering or community structure.

DOI: 10.1103/PhysRevE.101.062302

I. INTRODUCTION

Networks are a popular tool for representing and analyzing real-world systems consisting of entities and their relationships. They provide a simple yet intuitive representation for many complex systems. In the most basic incarnation, networks are simple graphs—undirected and unweighted with only one type of node and one type of edge. The usual picture is that nodes represent some group of objects (people, neurons, proteins, etc.), and edges represent some kind of interaction between them (friendship, synapses, binding, etc.) [1–8].

In many real-world settings, interactions are indicated by real-valued data, and so creating a simple network requires thresholding, which may take several forms [4–13]. The most obvious case of thresholding is when a continuous valued data set is explicitly thresholded by deciding what level of interaction is sufficiently strong to count as an edge in the network. A more subtle case is that of experimental limitation: interactions that exist but are very weak or rare may not be observed. Even for binary valued data sets, the sampling method may hide an implicit thresholding mechanism. For example, one commonly uses a combination of a yeast-two-hybrid screen and biochemical assays to detect and verify edges in protein-protein interaction networks. These methods typically do not detect weak protein-protein interactions [14] and are thus equivalent to applying a threshold on the edge strength in protein-protein interaction networks. For another example, consider friendship networks. Most everyday interactions between people are presumably not strong enough to constitute friendship. At what point does a casual acquaintance cross over to the category of friend? When people list their friends, in a survey for instance, they will implicitly apply some criteria to filter the friends from the acquaintances. Nevertheless, an understanding of the properties one should expect to observe from thresholded relational data is currently lacking.

In this paper, we examine the properties of networks created by thresholding relational data. To do this, we introduce a basic model of the underlying relational data, which is then thresholded to produce edges in the network. The model is derived from three assumptions:

(i) All nodes are statistically identical.

(ii) Any correlations are local.

(iii) The underlying relational data are normally distributed.

All three of these assumptions—which are no doubt violated in real-world systems—are quite natural for a null model. Assumption (i), that all nodes are identical, severely constrains what correlation structures are admissible. In fact, only two free parameters remain in the covariance matrix once this assumption is made: a local correlation strength between edges that share nodes, and a global correlation strength between edges that do not share nodes. Assumption (ii) sets

[*] gcant@umich.edu

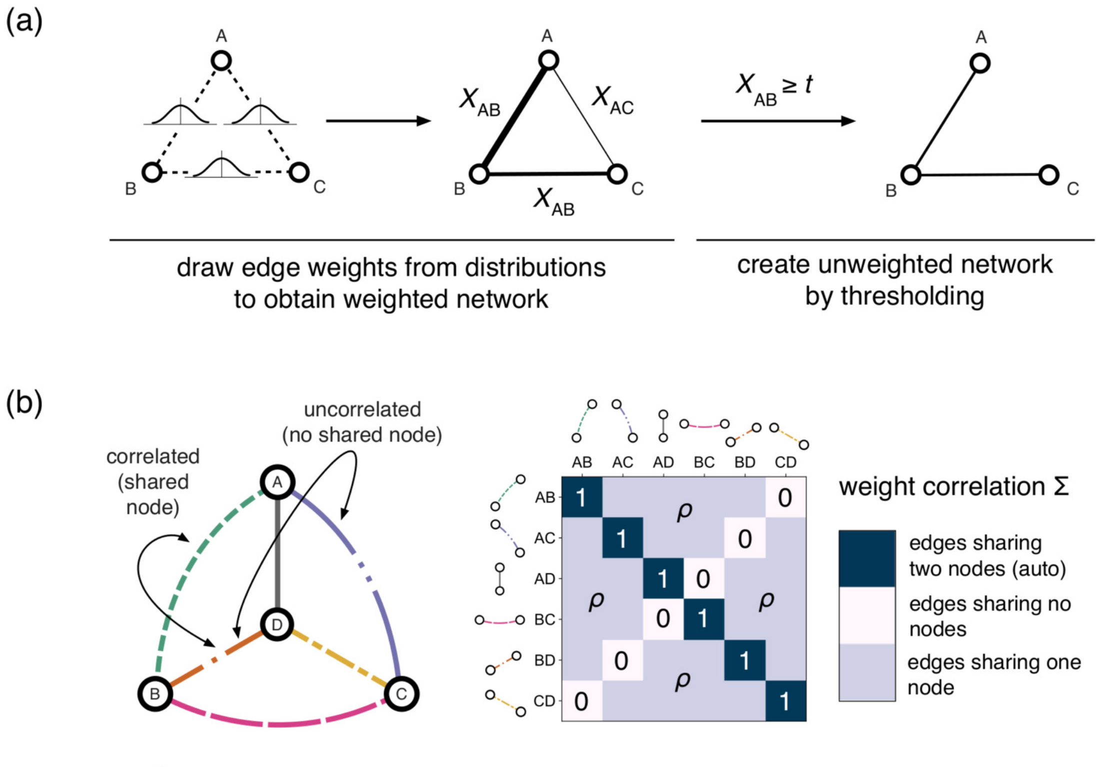
(a)
A
B
C
X_{AB}
X_{AC}
X_{AB}
$X_{AB} \geq t$
draw edge weights from distributions to obtain weighted network
create unweighted network by thresholding
(b)
correlated (shared node)
uncorrelated (no shared node)
A
B
C
D
AB
AC
AD
BC
BD
CD
1
0
ρ
weight correlation Σ
edges sharing two nodes (auto)
edges sharing no nodes
edges sharing one node

FIG. 1. Thresholding relational data to obtain networks. Panel (a) shows a general procedure to obtain unweighted networks from edge weights. Each edge weight is hypothesized to have been drawn from a specific distribution, generating an undirected weighted network. An unweighted network is then produced by assigning an edge whenever an edge weight X_{ij} is greater than a threshold t. In panel (b) we show how edge weights are correlated in the model of Sec. II by covariance matrix $\boldsymbol{\Sigma}$ [Eq. (5)]. Edge weights for edges that connect through a node have covariance $\mathrm{Cov}[X_{ij}X_{ik}] = \rho$, while edge weights not connected by a node have zero covariance.

the second of these to zero—edges that do not have a node in common are uncorrelated. The other free parameter, the local correlation strength, we call ρ. Our remaining freedom is to pick a distribution that is consistent with the required correlation matrix. The most obvious and simple choice is assumption (iii), the multivariate normal (Gaussian) distribution. We believe this to be the simplest nontrivial model for relational data.

The thresholding procedure will also be very simple: any of the relational data that fall above some threshold, t, will be said to constitute an edge in the network, and any that fall below will not. The threshold value t is a parameter of the model.

Sophisticated methods to extract networks from weighted data have been developed, for example in [11–13]. These more complicated methods lead to different networks, but in this paper we do not consider the relative merits of more advanced procedures. We favor the simplistic approach since it contains only one parameter and allows us to derive equations for several network properties. Nevertheless, in Appendix E we present some similar results for the so-called disparity filter of [11].

Our network ensemble on n nodes is thus defined by two parameters: the threshold, t, and a local correlation coefficient, ρ. Despite the simplicity of the model—the underlying relational data are normally distributed—we nonetheless find a number of the behaviors typically observed in complex networks, such as heavy-tailed degree distributions, short average path lengths, and large numbers of triangles. It does not, however, yield nonvanishing clustering or community structure in the large n limit, and so it cannot account for this observation in real-world data sets. Finally, the model we study is not constrained to produce positive-definite matrices. As a result, it is not immediately applicable to the study of thresholded correlation matrices (as, for example, in [10]).

This paper has two main parts. In Sec. II we define and justify the network model. Then, in Sec. III, we study the properties of the network ensemble. We look at the density of edges, triangles and clustering, the degree distributions, the shortest path lengths, and the giant component.

II. MODEL SPECIFICATION

A. Thresholding locally correlated data

A network can be represented by its adjacency matrix, $\boldsymbol{A}$, where $A_{ij} = 1$ if nodes i and j are connected and $A_{ij} = 0$ otherwise. We consider networks created by thresholding underlying relational data, $\boldsymbol{X}$, adding an edge between i and j if

$$X_{ij} \geqslant t. \tag{1}$$

To fully specify the model, we need to pick a distribution for $\boldsymbol{X}$ [see Fig. 1(a)]. Assuming that all nodes are statistically identical—exchangeable in the parlance of statistics—constrains our choice of distribution.

If nodes are identical, then the marginal distribution for X_{ij} must be the same for all (distinct) pairs i and j. Further, by a linear transform we can always set $E[X_{ij}] = 0$ and $\mathrm{Var}[X_{ij}] = 1$. So long as the appropriate transformation is made to t, this shift will have no effect on the thresholded network. For this reason, we will always assume X_{ij} has mean 0 and variance 1. Exchangeability puts further constraints on the covariance matrix, whose entries can take only three values. For i, j, k, l all distinct, these are

$$\begin{aligned}\mathrm{Var}\,[X_{ij}] &= \Sigma_{(i,j),(i,j)} = 1,\\ \mathrm{Cov}\,[X_{ij}, X_{ik}] &= \Sigma_{(i,j),(i,k)} = \rho,\\ \mathrm{Cov}\,[X_{ij}, X_{kl}] &= \Sigma_{(i,j),(k,l)} = \gamma,\end{aligned} \tag{2}$$

where $\mathrm{Cov}[X, Y]$ denotes covariance. We will assume that $\gamma = 0$ since this quantifies the correlation between two edges that do not share a node, i.e. two edges that do not "touch" [see Fig. 1(b)]. This leaves us with two free parameters, t and ρ. The remaining task is to pick a distribution with the required covariance matrix, $\boldsymbol{\Sigma}$.

In principle any distribution could be used, but the obvious choice is a multivariate normal distribution. In standard notation, a multivariate normal distribution (MVN) is denoted $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The probability density function of an N-dimensional MVN is

$$P(\boldsymbol{x}) = \frac{e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}}{\sqrt{(2\pi)^N \det(\boldsymbol{\Sigma})}}. \tag{3}$$

The normal distribution has many points in its favor. Famously, it arises in the central limit theorem, which makes it a plausible model for many random processes. If the relational data $\boldsymbol{X}$ arise due to the aggregation of many independent processes, then the central limit theorem implies $\boldsymbol{X}$ will be multivariate normally distributed. Further, the normal distribution is the maximum entropy distribution with the required covariance matrix, Eq. (2), and so could be justified as the "least informative distribution"—the model that makes the fewest extra assumptions beyond the correlation structure. We can also appeal to simple pragmatism: the multivariate normal distribution is well-studied and has convenient mathematical properties.

A concise statement of the model is as follows: given the freely chosen parameters $t \in \mathbb{R}$, $\rho \in [0, \frac{1}{2}]$, and the number of nodes n, draw a random variable $\boldsymbol{X}$ with

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{4}$$

where

$$\begin{aligned}\Sigma_{(i,j),(i,j)} &= 1,\\ \Sigma_{(i,j),(i,k)} &= \rho,\\ \Sigma_{(i,j),(k,l)} &= 0.\end{aligned} \tag{5}$$

Then create the network by thresholding $\boldsymbol{X}$,

$$A_{ij} = \begin{cases} 1 & \text{if } X_{ij} \geqslant t,\\ 0 & \text{otherwise.}\end{cases} \tag{6}$$

Note, we constrain $0 \leqslant \rho \leqslant \frac{1}{2}$ so that $\boldsymbol{\Sigma}$ is positive-semidefinite [15].

Even if we have good reason to believe that the marginal distributions for X_{ij} are not normal, the model may still be applicable. Consider an arbitrary cumulative distribution function, $F(x)$, and let $\Phi(x)$ denote the standard normal cumulative distribution function. If we sample $\boldsymbol{X}$ from a multivariate normal distribution, and then apply the function $F^{-1}(\Phi(x))$ to each X_{ij}, we will have transformed the edge weights to the arbitrary distribution F. So long as we apply the same transformation to t, however, the resulting network after thresholding will be identical.

The upshot is that our model can be adapted for any marginal distribution, and no network properties change—the assumption that the edge weights have normally distributed marginals is of no real consequence. What *is* important is the assumption that there is some transformation of the data such that the *joint* distribution is multivariate normal. While this assumption is a limitation, the above procedure is actually one of the standard methods for creating multivariate distributions with arbitrary marginals.

B. Sampling from the model

We now describe a simple algorithm to sample from the model. This algorithm also provides an intuitive model interpretation.

Let Z_i be n i.i.d. variables, $\mathcal{N}(0, 1)$. Let Y_{ij} be $\binom{n}{2}$ i.i.d. variables, $\mathcal{N}(0, 1)$. Then let

$$W_{ij} = \sqrt{1-2\rho}\,Y_{ij} + \sqrt{\rho}(Z_i + Z_j). \tag{7}$$

Note that W_{ij} is normally distributed with mean zero, and further

$$\begin{aligned}\mathrm{Var}\,[W_{ij}] &= 1,\\ \mathrm{Cov}\,[W_{ij}, W_{ik}] &= \rho,\\ \mathrm{Cov}\,[W_{ij}, W_{kl}] &= 0.\end{aligned} \tag{8}$$

Hence, $\boldsymbol{W}$ is distributed identically to $\boldsymbol{X}$. So, to sample from the model:

(i) Sample $\boldsymbol{z}$, a length n-vector of i.i.d. standard normal variables.

(ii) For $i < j$, generate $y \sim \mathcal{N}(0, 1)$, and if

$$y > \frac{t - \sqrt{\rho}(z_i + z_j)}{\sqrt{1-2\rho}} \tag{9}$$

add edge (i, j) to the network.

If $\rho = \frac{1}{2}$, generating y is unnecessary and one can simply add edge (i, j) if $\sqrt{1/2}(z_i + z_j) \geqslant t$.

A Python package to generate networks along with scripts for the figures in this paper is publicly available [16].

To achieve the required correlations, the algorithm above separates X_{ij} into node and edge effects. Each node is given a value Z_i, and X_{ij} is created by a linear combination of Z_i and Z_j plus i.i.d. random noise Y_{ij}. We can interpret the Z's as latent variables that control the propensity for individual nodes to have edges, and ρ controls the relative strength of the noise process. When $\rho = 1/2$ edges are entirely determined by the values of Z, while at $\rho = 0$ edges are entirely random and independent.

Despite this equivalent formulation, our model should not be primarily understood as a latent variable model since it was not constructed as one. Rather, the equivalent latent variable model is derived and used for algorithmic convenience.

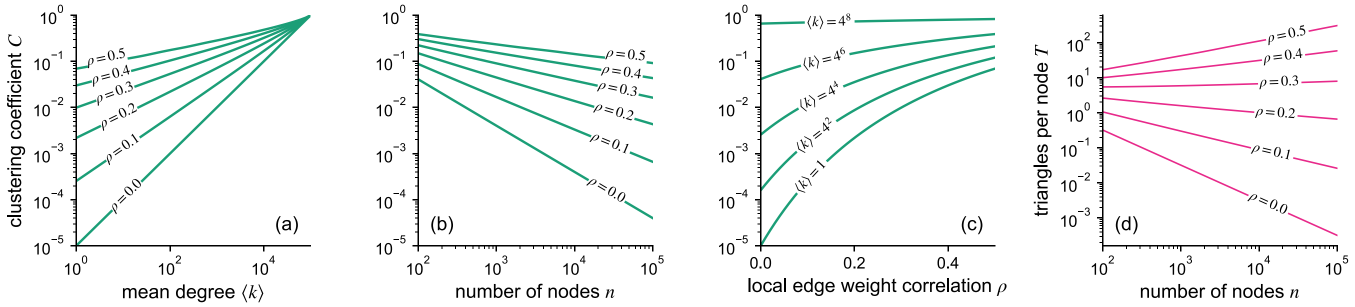

FIG. 2. Clustering C and triangles per node T as computed in Sec. III B. Clustering decreases with increasing number of nodes, however the number of triangles per node increases with a growing number of nodes n for large values of ρ. Clustering increases both with increasing mean degree $\langle k\rangle$ and local edge weight correlation ρ. In panels (a) and (c) we chose $n = 100\,000$ and in panels (b) and (d) we fixed $\langle k\rangle = 4$.

In fact, the existence of this latent variable interpretation is not surprising. As $n \to \infty$, our model is in a class of models known as *exchangeable random graphs* [17,18]. The Aldous-Hoover theorem implies that all exchangeable random graphs have an equivalent latent variable model [17–19].

III. NETWORK PROPERTIES

We now turn our attention to the properties of the networks created by the model.

A. Edge density

Edges in the network exist whenever the corresponding weight X_{ij} is greater than t. The marginal distribution for X_{ij} is simply a standard normal distribution. Thus,

$$E[A_{ij}] = P[A_{ij} = 1] = P[X_{ij} \geqslant t] = 1 - \Phi(t), \qquad (10)$$

where $\Phi(x)$ is the cumulative distribution function for the standard normal distribution $\mathcal{N}(0, 1)$. When $\rho = 0$, all edges exist independently and the model is equivalent to the random graph, $G_{n,p}$, with $p = 1 - \Phi(t)$.

The mean degree is equally simple to compute. For all ρ,

$$E[k_i] = \sum_{j\neq i} E[A_{ij}] = (n-1)[1 - \Phi(t)]. \qquad (11)$$

If we want to pick t for a desired mean degree $\langle k\rangle$, it is easy to invert this to obtain

$$t = \Phi^{-1}\left(1 - \frac{\langle k\rangle}{n-1}\right). \qquad (12)$$

B. Triangles, clustering, and degree variance

Many complex networks are observed to have large numbers of triangles. The clustering coefficient or transitivity is one way to quantify this. We can quantify the clustering with the probability that a triangle is closed, given that two of its edges already exist,

$$C = P[A_{ik} = 1 | A_{ij}, A_{jk} = 1] = \frac{P[A_{ik}, A_{ij}, A_{jk} = 1]}{P[A_{ij}, A_{jk} = 1]}. \qquad (13)$$

The numerator of this equation corresponds to the density of triangles while the denominator corresponds to the density of two-stars (which also determines the variance of the degree distribution). Note that for simplicity we shorten the logical connective “and” (or “$\wedge$”) using commas, e.g., $P[A_{ij} = 1 \,\wedge\, A_{jk} = 1] \equiv P[A_{ij}, A_{jk} = 1]$.

The marginal distributions of a MVN are themselves MVN, and they are found by simply dropping the unwanted rows and columns in the correlation matrix $\boldsymbol{\Sigma}$. Thus, $(X_{ij}, X_{ik})^T$ will be bivariate normally distributed and $(X_{ij}, X_{ik}, X_{jk})^T$ will be trivariate normally distributed, both with correlation coefficient ρ. Introducing the Hermite polynomials $H_N(x)$ as defined in Appendix A, one finds that

$$P[X_{ij}, X_{ik} \geqslant t] = \sum_{N=0}^{\infty} \frac{\rho^N}{N!}[\phi(t)H_{N-1}(t)]^2 \qquad (14)$$

for the density of two-stars and

$$P[X_{ij}, X_{ik}, X_{jk} \geqslant t] = \sum_{N=0}^{\infty}\sum_{i=0}^{N}\sum_{j=0}^{N-i} \frac{\rho^N \phi(t)^3}{i!\, j!\, (N-i-j)!} \times H_{N-1-i}(t)H_{N-1-j}(t)H_{i+j-1}(t) \qquad (15)$$

for triangles. Both sums converge for $\rho \leqslant 0.5$, and we can estimate them accurately with a finite number of terms [20]. Noting that there are $\binom{n-1}{2}$ potential triangles for each node, the expected number of triangles per node is simply $\binom{n-1}{2}$ times their density

$$T = \binom{n-1}{2} P[X_{ij}, X_{ik}, X_{jk} \geqslant t]. \qquad (16)$$

Plots of these functions are shown in Fig. 2. We find that T is much larger in these networks than in the random graph $G_{n,p}$—larger by multiple orders of magnitude. In fact, while T goes to zero in the large n limit for the random graph, in this model we find that T increases with n for large values of ρ. On the other hand, the clustering coefficient C decreases with a growing number of nodes for all parameter values. This leads to a slightly paradoxical result for large ρ: in the limit $n \to \infty$, the expected number of triangles at each node goes to infinity, and the clustering coefficient still goes to zero. The reason for this is that the number of two-stars diverges faster than the number of triangles.

Equation (14) can also be used to compute the variance of the degree distribution. To see this, note that a node of degree k has $\binom{k}{2}$ two-stars. Further, noting that there are $\binom{n-1}{2}$ potential

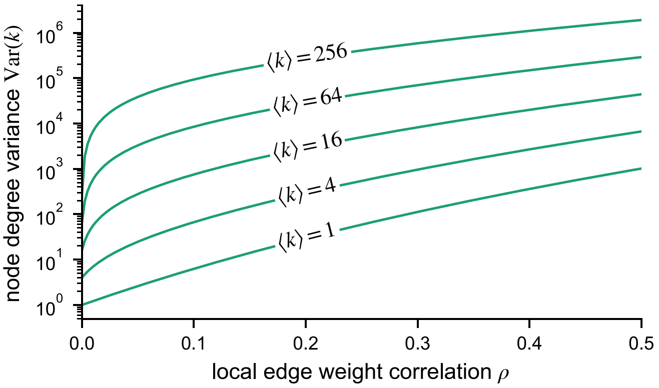

FIG. 3. The variance of degree, Eq. (18), increases with ρ, the local edge weight correlation. With increasing mean degree $\langle k\rangle$, even small correlations ρ produce networks of significantly broader degree distribution than the random graph $G_{n,p}$.

two-stars (the same number of potential triangles), we find

$$\frac{1}{2}(\langle k^2\rangle - \langle k\rangle) = \binom{n-1}{2} P[X_{ij}, X_{ik} \geqslant t]. \tag{17}$$

Combining this with Eq. (11), the variance of the node degree k can be written

$$\begin{aligned}\mathrm{Var}[k] &= (n-1)\Phi(t)[1-\Phi(t)] \\ &\quad + (n-1)(n-2)\sum_{N=1}^{\infty}\frac{\rho^N}{N!}[\phi(t)H_{N-1}(t)]^2. \end{aligned}\tag{18}$$

The first term is simply the variance of a binomial distribution. For $\rho = 0$ the second term vanishes and we recover the correct result for the random graph $G_{n,p}$. For $\rho > 0$ the sum is positive and monotonically increases with ρ as illustrated in Fig. 3.

C. Degree distribution

In the previous two subsections, we gave expressions for the mean and variance of the degrees. Here we give expressions for the full distribution of degrees.

The degree distribution p_k is the probability that a node has k edges. For this model, the degree distribution can be written

$$p_k = \binom{n-1}{k}\sqrt{\frac{1-\rho}{2\pi\rho}}\int_{-\infty}^{\infty} e^{f_k(y)}dy, \tag{19}$$

where

$$\begin{aligned} f_k(y) &= k\ln[1-\Phi(y)] + (n-k-1)\ln[\Phi(y)] \\ &\quad - \frac{1}{2}\left(\frac{t-\sqrt{1-\rho}y}{\sqrt{\rho}}\right)^2. \end{aligned}\tag{20}$$

This result is derived in Appendix B.

The integral in Eq. (19) can be computed numerically to high precision using Gauss-Hermite quadrature, centered at the maximum of $f_k(y)$. Increasing the order of Gauss-Hermite quadrature (i.e., incorporating more points) increases the accuracy. The full details are in Appendix B.

We can also approximate the integral using Laplace's method [21], an asymptotic approximation for integrals of this form (equivalent to a first order Gauss-Hermite quadrature). The idea of the method is to replace the function $f_k(y)$ by a second-order Taylor series around its maximum. For large n, the last (quadratic) term in f_k will be negligible, and for $0 < k < n-1$ the maximum will be at

$$y_{0,k} = \Phi^{-1}\left(1 - \frac{k}{n-1}\right). \tag{21}$$

Combining this with Stirling's approximation for the binomial coefficient, we find

$$\begin{aligned} p_k \sim \frac{1}{n-1}\sqrt{\frac{1-\rho}{\rho}}\exp\bigg[&-\left(\frac{1-2\rho}{2\rho}\right)y_{0,k}^2 \\ &+ \left(\frac{t\sqrt{1-\rho}}{\rho}\right)y_{0,k} - \frac{t^2}{2\rho}\bigg]. \end{aligned}\tag{22}$$

Together with the closed-form approximation for Φ^{-1}, given in Appendix C, Eq. (22) provides a closed-form approximation for the degree distribution.

Figure 4 shows some example degree distributions, computed to high precision using Eq. (19) along with the asymptotic approximation, Eq. (22), where we chose $n = 100\,000$ and $\langle k\rangle = 100$.

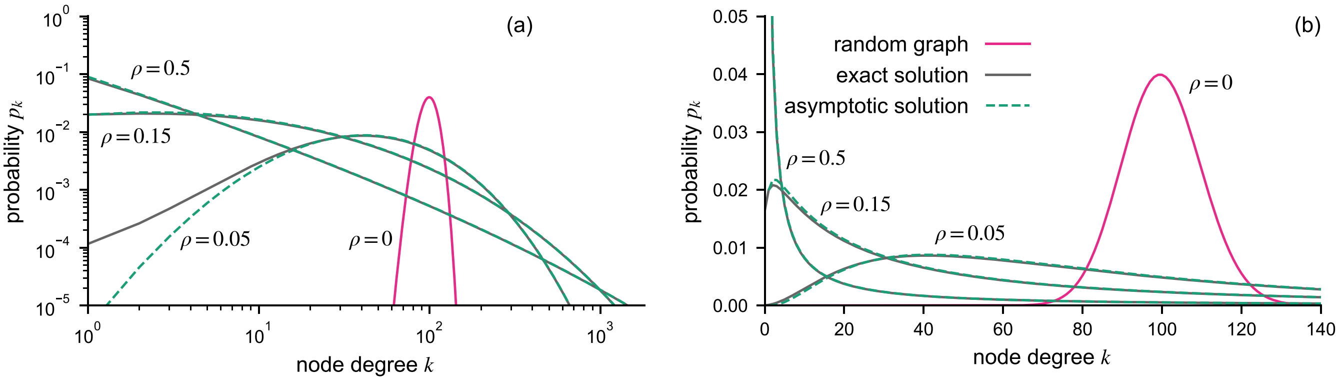

FIG. 4. We show degree distributions computed using Eq. (19) for $n = 100\,000$ and $\langle k\rangle = 100$ for increasing local edge weight correlation ρ in log-log (a) and linear scales (b). We also compare them to the asymptotic approximation Eq. (22). Note that large values of ρ produce broad degree distributions, which could be easily mistaken for log-normal or power-law distributions.

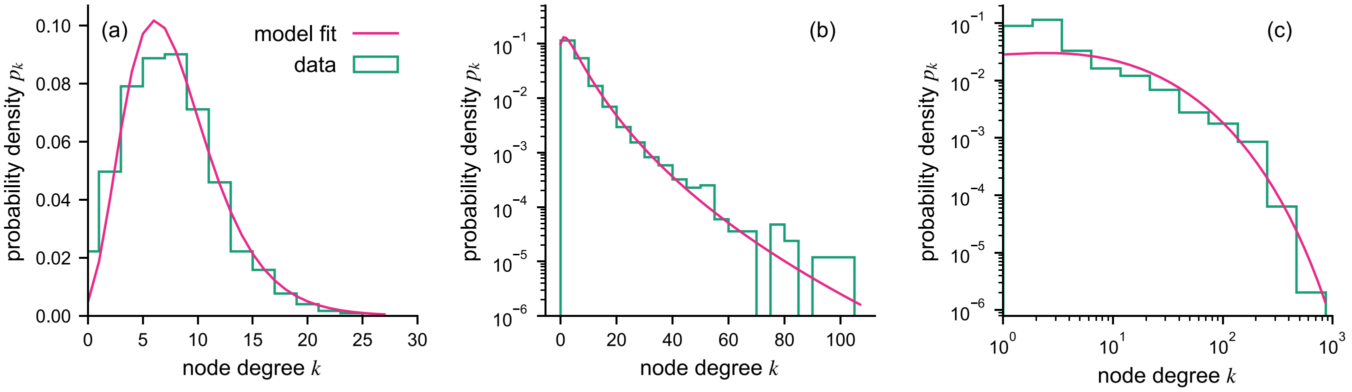

FIG. 5. Degree histograms for the three real-world networks introduced in Sec. III C along with fitted distributions from the thresholded normal model. We show (a) a high school friendship network, (b) a coauthorship network between scientists, and (c) a protein-protein interaction network.

To illustrate how these degree distributions compare to the degree distributions of real networks, we chose three data sets from different domains, and fit the model. The first data set is a network of friendships between students at a U.S. high school ($n = 2587$) [22], the second data set is a co-authorship network of researchers ($n = 16\,726$) [23], and the third network describes interactions between proteins ($n = 6327$) [24].

Given a number of nodes n, the model studied in this paper has two free parameters, t and ρ. A simple procedure to fit the model to the data is to choose t and ρ so that the mean and variance of the model's degree distribution match the observed values. We use Eq. (12) to fix t and subsequently Newton's method to solve Eq. (18) for ρ.

The results of this exercise are shown in Fig. 5. The networks were chosen for their different degree distributions—note the different scales on the axes: linear, log-linear, and log-log—and the threshold model can qualitatively ape these distributions. Nevertheless, the similarity of the degree distribution should not be overemphasized. As discussed, this model has vanishing clustering, so it cannot account for this observation of real-world networks. Comparisons for clustering are shown in Appendix D.

While the degrees in the thresholded networks, $k_i = \sum_j A_{ij}$, in general follow a complicated distribution, the underlying degrees $d_i = \sum_j X_{ij}$ are always normally distributed. When $\rho = 0$, d_i is Gaussian and k_i is binomial, or Poisson in the sparse limit. When $\rho > 0$, d_i is still Gaussian, but k_i now follows a heavy-tailed distribution. Thus, the heavy-tailed distribution observed in the model is due to the combination of correlation and thresholding. Without positive correlation we observe Poisson distributions; without thresholding we observe Gaussian distributions.

D. Giant component

A well-studied problem in the theory of random graphs is the formation of a large connected (giant) component. At very low densities, only a handful of nodes can be reached from any other node, but at some critical point a macroscopic number of nodes will be connected. For the random graph this transition occurs at a mean degree of $\langle k\rangle = 1$ [1,25,26].

To explore the effects of $\rho > 0$ we sampled from the model as described in Sec. II B and measured the size of the second largest component as a susceptibility parameter for the phase transition. The maximum of this susceptibility parameter is used to find the transition lines in Fig. 6(a).

We find that as ρ or n increases, the transition occurs at lower values of the mean degree. This result is in line with the configuration model for which the transition point decreases with increasing variance in the degree distribution. For $\rho = 0$ we recover the standard result for the random graph.

For the other limit case, $\rho = 1/2$, recall that all edge weights can be considered to arise from node "propensities," Z_i, with $X_{ij} = \sqrt{1/2}(Z_i + Z_j)$. This implies that all nodes that are connected to any other nodes must also be connected to the node with maximum propensity $Z_{\max}$. The size of the largest component is then given by this node's degree plus 1, $k_{\max} + 1$. The second largest component is then always of size 1. We therefore omit $\rho = 1/2$ in the numerical analysis.

E. Shortest path lengths

Another phenomenon well established in the complex networks literature is that randomly chosen nodes often have surprisingly short paths between them. This is often referred to as the "six degrees of separation" or "small-world" phenomenon [1,27]. By a common definition, network models are considered to demonstrate this property if the average shortest path length $\langle d_{ij}\rangle$ between nodes grows logarithmically (or slower) as the number of nodes increases [1].

Using the method described in Sec. II B, we sampled from the threshold model to verify that it displays this property. We looked at networks with between 100 and 30 000 nodes, with mean degree $\langle k\rangle = 5$, and we investigated the influence of increasing edge weight correlation ρ. After sampling a network from the model, we computed the average shortest path length $\langle d_{ij}\rangle$ on the largest (giant) component. For each parameter combination, we computed the mean by averaging 200 sampled networks.

The results are shown in Fig. 7(a). Since it is well known that the random graph $G_{n,p}$ has short shortest paths [28], it is unsurprising that the threshold model does also (recall, for $\rho = 0$ they are equivalent, and we see the standard $\langle d_{ij}\rangle \propto \log n$ scaling behavior). For $\rho > 0$, we see that the average

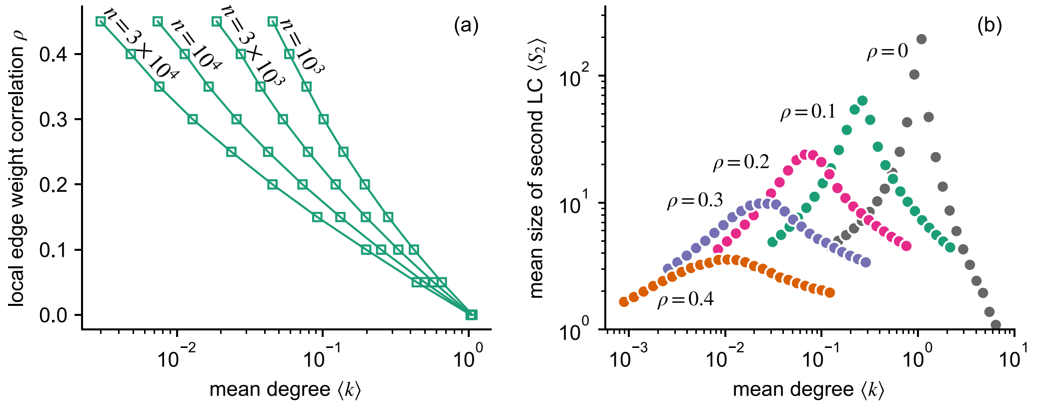

FIG. 6. Simulations with $1000 \leqslant n \leqslant 30\,000$, mean degree $10^{-3} \leqslant \langle k\rangle \leqslant 10$, $0 \leqslant \rho \leqslant 0.45$. 1000 samples were taken for each of the parameter combinations. Panel (a) shows the points of transitions for an increasing number of nodes n. To the left of the line, the network does not possess a giant component, while to the right it does. The transition point was computed using the mean size of the second largest component as a susceptibility parameter. Panel (b) shows an example of the susceptibility parameter for $n = 10\,000$.

shortest path lengths grow significantly slower than logarithmically, a behavior sometimes referred to as "ultra-small-world" and often related to networks with power-law degree distribution [29,30]. In our model, the effect appears despite the fact that the degree distribution does not follow a power law.

As discussed, when $\rho = 1/2$ all edge weights can be considered to arise from node propensities Z_i, such that $X_{ij} = \sqrt{1/2}(Z_i + Z_j)$. All nodes are then either disconnected or part of the giant component, and the node with maximum propensity $Z_{\max}$ is connected to all nodes in the giant component. Hence, all nodes in the giant component are either directly connected or can reach each other in two steps through the maximum-degree node. So, when $\rho = 1/2$ the average shortest path length must be $1 \leqslant \langle d_{ij}\rangle < 2$.

IV. DISCUSSION

In this paper, we studied the effects of thresholding relational data. We started with a simple model of multivariate normally distributed data, with only one free parameter, ρ, controlling local correlations. We then demonstrated that thresholding this normally distributed correlated relational data reproduces many of the properties commonly associated with complex networks. In particular, we find that the combined effects of correlation and thresholding lead to heavy-tailed degree distributions, relatively large numbers of triangles, and short average path lengths.

The underlying relational data $\boldsymbol{X}$ in the model we introduce would not usually be considered complex. It is generated from a highly symmetric multivariate normal distribution with only one free parameter. Since every pair of nodes has some level of interaction, the graphical interpretation for $\boldsymbol{X}$ would be a weighted complete graph, with all edge weights (and linear combinations thereof) normally distributed. For example, the "degrees," $d_i = \sum_j X_{ij}$, are normally distributed. And yet, after thresholding the networks show several properties commonly associated with complex networks.

One way to think about these results is in the context of the central limit theorem. Whenever interaction strengths are the aggregate result of a large number of processes, then we expect $\boldsymbol{X}$ to be normally distributed. Constructing a simple

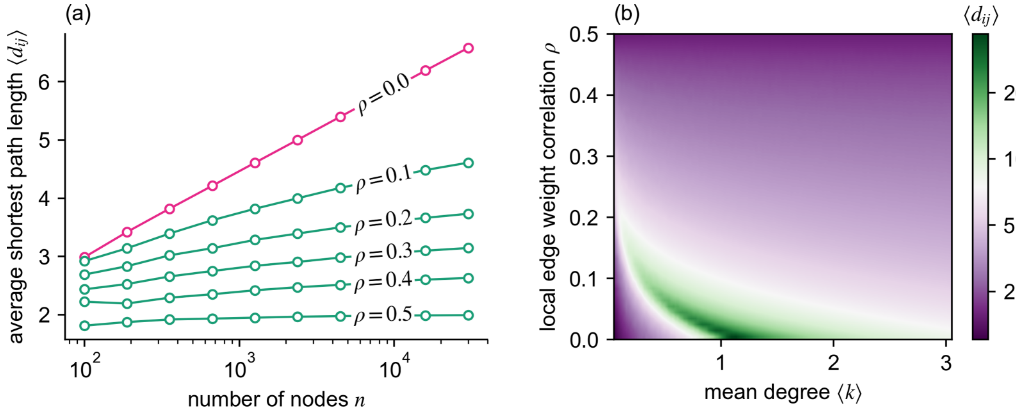

FIG. 7. Panel (a) shows the scaling of the average shortest path in the largest connected component with the number of nodes, n. We fix the mean degree $\langle k\rangle = 5$, and each point is averaged over 200 samples. For $\rho = 0$ we recover the result for the random graph $G_{n,p}$, where $\langle d_{ij}\rangle \propto \log n$. For nonzero correlation, the average shortest path length increases slower than logarithmically. In panel (b) we show the average shortest path length for different mean degrees and values of ρ for networks with $n = 10\,000$, again sampled 200 times for each parameter combination.

graph from these data can lead to complex networks. This provides one simple explanation for the ubiquity of complex networks—they can arise as a consequence of the central limit theorem.

Of course, for most scientific questions of interest the exact details of the mechanisms and structure are what matter. In a social network, for example, answering the question "who influences whom, and why?" is far from trivial, and the fact that the network has certain commonly observed properties is usually incidental.

In summary, straightforward assumptions lead to several of the properties associated with complex networks. If a network arises by a simple thresholding procedure, then finding that it is "complex" need be no more surprising than finding a bell-shaped curve in a regular data set.

ACKNOWLEDGMENTS

The authors thank the students, staff, and faculty of the Santa Fe Institute Complex Systems Summer School for bringing them together and providing a stimulating research environment. All authors made significant contributions to the project. G.T.C. and B.F.M. wrote the final manuscript. B.F.M. was supported as an Add-On Fellow for Interdisciplinary Life Science by the Joachim Herz Stiftung. A.C.S. was supported by the Clarendon Fund, e-Therapeutics plc, and the Engineering and Physical Sciences Research Council under Grant No. EP/L016044/1. J.S. was supported by the U.S. Army Research Office under Multidisciplinary University Research Initiative Award No. W911NF-13-1-0340. G.S. was supported by the Natural Sciences and Engineering Research Council of Canada and the Sentinel North program.

APPENDIX A: MULTIVARIATE NORMAL INTEGRALS AND HERMITE POLYNOMIALS

The probability of a two-star existing with nodes i, j, and k as constituents is given by

$$P[X_{ij}, X_{ik} \geqslant t] = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_t^\infty \int_t^\infty e^{-\frac{1}{2}\left(\frac{x^2-2\rho xy+y^2}{1-\rho^2}\right)} dx\, dy. \tag{A1}$$

Direct computation of the integral is not straightforward, but we can compute it quickly using the Hermite polynomials [20]. A quick outline of this method is as follows: for $n \geqslant 0$, define the Hermite polynomials as

$$H_n(x) = (-1)^n e^{\frac{x^2}{2}} \frac{d^n}{dx^n} e^{-\frac{x^2}{2}}. \tag{A2}$$

As the name suggests, the Hermite polynomials are in fact polynomials, for example $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = x^2 - 1$, and so on. For notational convenience, also define

$$H_{-1}(x) = \frac{1-\Phi(x)}{\phi(x)}. \tag{A3}$$

Using the Hermite polynomials, we can expand Eq. (A1) as an infinite sum and integrate term by term. The final result is given by Eq. (14). The same trick is used for the three-dimensional integral to give Eq. (15).

APPENDIX B: DEGREE DISTRIBUTION

Since, by assumption, all nodes in this model are equivalent, we will simply consider the one-node marginal to compute the degree distribution. Let $\boldsymbol{U}$ be all the terms in $\boldsymbol{X}$ that are associated with node 0, i.e., $U_j = X_{0j}$. Then, $\boldsymbol{U}$ is multivariate normally distributed, $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^{(0)})$, where $\boldsymbol{\Sigma}^{(0)}$ has ones along the diagonal and ρ everywhere else,

$$\Sigma^{(0)}_{jk} = \Sigma_{(0,j),(0,k)} = \begin{cases} 1 & \text{for } j = k, \\ \rho & \text{otherwise.} \end{cases}$$

The focal node will have degree k when exactly k terms in $\boldsymbol{U}$ are larger than the threshold t. There are $\binom{n-1}{k}$ different ways this can happen, and each is equally likely. So, to compute p_k we can compute the probability that the first k terms in $\boldsymbol{U}$ are larger than t and all others are smaller, and then multiply by $\binom{n-1}{k}$ to obtain

$$p_k = \binom{n-1}{k} P[U_1, \dots, U_k \geqslant t; U_{k+1}, \dots, U_{n-1} < t]. \tag{B1}$$

To solve this integral, we use a standard trick [31]. First, we note that if $Z_0, Z_1, \dots, Z_{n-1}$ are i.i.d. $\mathcal{N}(0, 1)$, then

$$((\sqrt{1-\rho}Z_1 + \sqrt{\rho}Z_0), \dots, (\sqrt{1-\rho}Z_{n-1} + \sqrt{\rho}Z_0))^T \tag{B2}$$

will be distributed identically to $\boldsymbol{U}$. Further, once we know the value of Z_0, then all the terms are independent, and the probability that any one of them is greater than t is the probability that $Z_1 \geqslant \frac{t-\sqrt{\rho}z}{\sqrt{1-\rho}}$. Given $Z_0 = z$, the probability that exactly k values will be greater than t and the rest less than t is

$$\binom{n-1}{k} \left[1 - \Phi\left(\frac{t-\sqrt{\rho}z}{\sqrt{1-\rho}}\right)\right]^k \Phi\left(\frac{t-\sqrt{\rho}z}{\sqrt{1-\rho}}\right)^{n-1-k}. \tag{B3}$$

Averaging this quantity over z then provides us with the correct expression,

$$p_k = \binom{n-1}{k} \times \underbrace{\int_{-\infty}^{+\infty} \left[1 - \Phi\left(\frac{t-\sqrt{\rho}z}{\sqrt{1-\rho}}\right)\right]^k \Phi\left(\frac{t-\sqrt{\rho}z}{\sqrt{1-\rho}}\right)^{n-1-k} \phi(z)dz}_{=I_{n,k}}, \tag{B4}$$

where $I_{n,k}$ is the integral. A change of variables allows us to write

$$I_{n,k} = \sqrt{\frac{1-\rho}{2\pi\rho}} \int_{-\infty}^{\infty} e^{f_k(y)} dy, \tag{B5}$$

where

$$f_k(y) = k \ln[1-\Phi(y)] + (n-k-1)\ln[\Phi(y)] - \frac{1}{2}\left(\frac{t-\sqrt{1-\rho}y}{\sqrt{\rho}}\right)^2. \tag{B6}$$

A standard approach to approximate such an integral is to use Laplace's method. In this approach, one expands f about

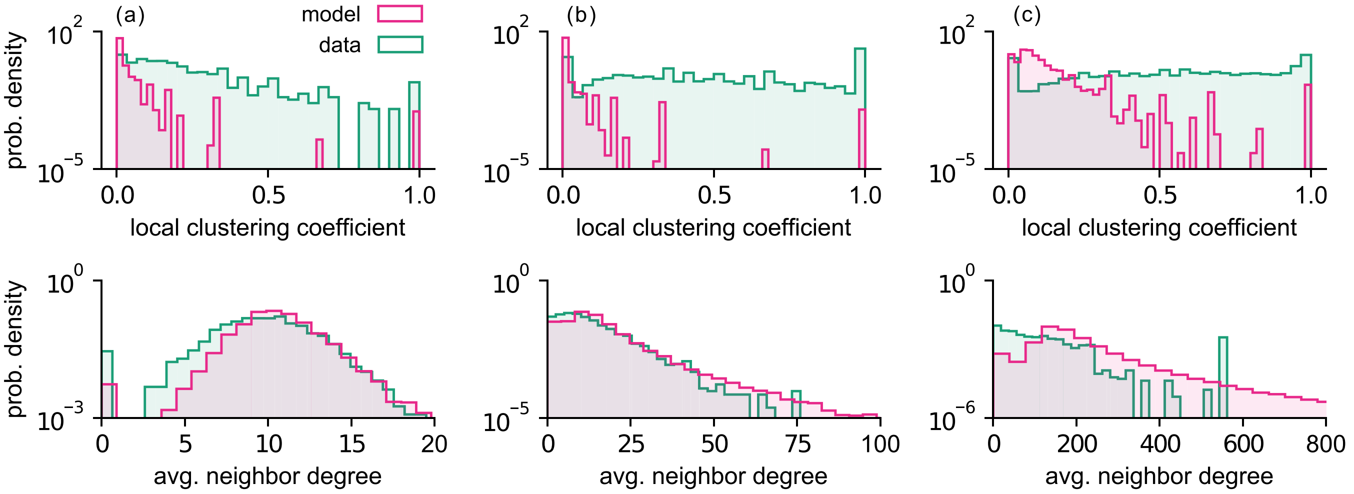

FIG. 8. Local clustering coefficient and average neighbor degree for the three real-world networks introduced in Sec. III C along with simulated results from the thresholded normal model. We show (a) a high school friendship network, (b) a co-authorship network between scientists, and (c) a protein–protein interaction network.

its maximum and then neglects higher-order terms,

$$f(y) \approx f(y_0) - \frac{|f''(y_0)|}{2}(y - y_0)^2.$$

Having done this, the integral reduces to a standard Gaussian integral. While this approach is asymptotically correct (in the large n and k limit), we can improve the approximation by including more terms using Gauss-Hermite quadrature. Rewriting the integral again, and making another change of variables:

$$I_{n,k} = \sqrt{\frac{1-\rho}{2\pi\rho|f_k''(y_0)|}}e^{f_k(y_0)}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}+R_k\left(\frac{x}{\sqrt{|f_k''(y_0)|}}+y_0\right)}dx, \tag{B7}$$

where R_k is the remaining terms of f_k after expansion:

$$R_k(y) = f_k(y) - f_k(y_0) + \frac{|f_k''(y_0)|}{2}(y - y_0)^2. \tag{B8}$$

Now we can approximate the integral using Gauss-Hermite quadrature:

$$I_{n,k}(N) = \sqrt{\frac{1-\rho}{2\pi\rho|f_k''(y_0)|}}e^{f_k(y_0)}\left[\sum_{i=1}^{N} w_i e^{R_k\left(\frac{x_i}{\sqrt{|f_k''(y_0)|}}+y_0\right)}\right], \tag{B9}$$

where x_i are the points for which $H_N(x_i) = 0$ and the weights w_i are

$$w_i = \frac{N!\sqrt{2\pi}}{N^2[H_{N-1}(x_i)]^2}. \tag{B10}$$

Note that $I_{n,k}(1)$ is Laplace's approximation, i.e., Laplace's approximation is a first-order Gauss-Hermite quadrature at the maximum of f_k, while $I_{n,k}(N)$ approximates the remainder terms with increasingly high order polynomials, and so we expect $I_{n,k}(N) \to I_{n,k}$ as N increases.

APPENDIX C: APPROXIMATION OF INVERSE CUMULATIVE DISTRIBUTION FUNCTION

The normal distribution's inverse cumulative distribution function, $\Phi^{-1}(x)$, can be approximated [32] for $0 < x \leqslant 0.5$ as

$$\Phi^{-1}(x) \approx \frac{a_0 + a_1 s}{1 + b_1 s + b_2 s^2} - s, \quad s = \sqrt{-2\ln(x)} \tag{C1}$$

with

$$a_0 = 2.307\,53, \quad b_1 = 0.992\,29, \tag{C2a}$$

$$a_1 = 0.270\,61, \quad b_2 = 0.044\,81. \tag{C2b}$$

For $0.5 < x \leqslant 1$ we use $\Phi^{-1}(x) = -\Phi^{-1}(1-x)$.

APPENDIX D: OTHER PROPERTIES OF REAL NETWORKS

In Fig. 8 we compare simulations from the thresholded normal model to the real networks from Sec. III C. Other degree properties, such as the average neighbor-degree, seem to be modeled well, while the local clustering coefficient is generally smaller in the simulations than real data, as expected.

APPENDIX E: DISPARITY FILTERING

The simplistic style of thresholding that we have considered is not the only method to extract a network from relational data. Indeed, more sophisticated algorithms have been developed [11–13]. By limiting our analysis to a simple thresholding procedure, rather than a more sophisticated algorithm, we have been able to derive several of the basic properties. In this Appendix, we repeat some of the analysis for the more complex "disparity filter" algorithm [11].

The disparity filter assumes that the relational data are positive and vary by orders of magnitude. In contrast, we have considered data that are distributed according to a standard normal distribution—X_{ij} is negative half of the time and will virtually never be larger in magnitude than 10. To match the assumptions of the disparity filter, we apply the algorithm to $e^{X_{ij}}$.

Since X_{ij} is normally distributed, $e^{X_{ij}}$ is log-normally distributed. And, since the exponential function is monotonic, a simple thresholding procedure at the value of e^t is equivalent to our previous analyses. However, since $e^{X_{ij}}$ is non-negative and will vary by orders of magnitude, these data also match the assumptions of the disparity filter.

The disparity filter algorithm proceeds by assigning a local significance score $\alpha_{ij} \neq \alpha_{ji}$ to each potential edge. In the normal model, this score will be

$$\alpha_{ij} = \left(\frac{\sum_{k\neq j} e^{X_{ik}}}{\sum_k e^{X_{ik}}}\right)^{n-2}. \tag{E1}$$

An edge is considered present between nodes i and j if either α_{ij} or α_{ji} exceeds a predetermined significance threshold, α.

Making use of the decomposition from Eq. (8), we can write

$$\alpha_{ij} = \left(\frac{\sum_{k\neq j} e^{\sqrt{1-2\rho}Y_{ik}+\sqrt{\rho}Z_k}}{\sum_k e^{\sqrt{1-2\rho}Y_{ik}+\sqrt{\rho}Z_k}}\right)^{n-2}. \tag{E2}$$

Further, we can make a mean-field style approximation and replace $\sum_{k\neq j} e^{\sqrt{1-2\rho}Y_{ik}+\sqrt{\rho}Z_k}$ by its expected value to arrive at

$$\alpha_{ij} = \left(\frac{(n-2)e^{\frac{1-\rho}{2}}}{(n-2)e^{\frac{1-\rho}{2}} + e^{\sqrt{1-2\rho}Y_{ij}+\sqrt{\rho}Z_j}}\right)^{n-2}. \tag{E3}$$

Defining the constant

$$c = \frac{2\log\left((n-2)(\alpha^{-\frac{1}{n-2}}-1)\right)+(1-\rho)}{2\sqrt{\rho}}, \tag{E4}$$

a quick calculation establishes that either α_{ij} or α_{ji} is larger than α with probability

$$E[A_{ij}] = \int_{-\infty}^{+\infty} (1-\Phi(-x\sqrt{1/\rho-2}+c)^2)\phi(x)dx, \tag{E5}$$

where ϕ and Φ are again the density function and cumulative distribution function for the standard normal distribution.

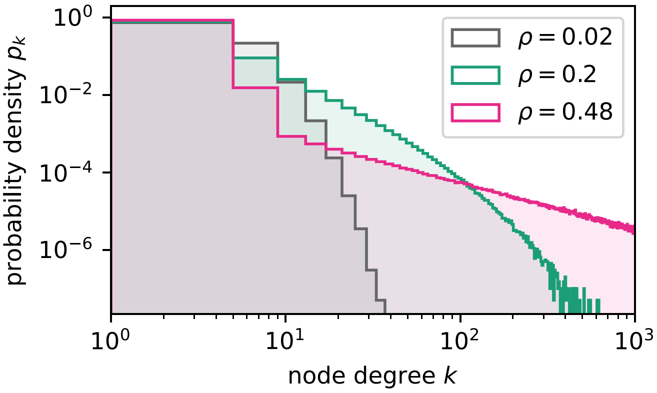

FIG. 9. Simulations for the disparity filter [11] applied to $e^{X_{ij}}$. Networks with 2,000 nodes were created by first sampling matrices $\boldsymbol{X}$ from a normal distribution and then applying the disparity filter to $e^{X_{ij}}$. Three different values for ρ were used and α was set using Eq. (E5) so that the final network would have a mean degree of 4.

This integral can easily be computed using Gauss-Hermite quadrature. Its derivatives with respect to α are also simple to compute, and so root finding algorithms such as Newton's method can find the correct choice of α for a desired final density.

While the above derivation for the network density in this model is reasonably straightforward, other properties such as the degree distribution are more challenging. Instead of re-deriving the results of this paper for a sophisticated filtering algorithm, we present the results of simulations in Fig. 9. Increasing the correlation ρ similarly increases the variance on the node degree k. This demonstrates that the qualitative properties derived for the naive thresholding still apply for at least some more sophisticated thresholding methods.

[1] M. Newman, *Networks: An Introduction*, 2nd ed. (Oxford University Press, Oxford, 2018).

[2] J. Berg, M. Lässig, and A. Wagner, Structure and evolution of protein interaction networks: A statistical model for link dynamics and gene duplications, BMC Evol. Biol. **4**, 51 (2004).

[3] E. T. Bullmore and D. S. Bassett, Brain graphs: graphical models of the human brain connectome, Ann. Rev. Clin. Psych. **7**, 113 (2011).

[4] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, J. E. Larsen, and S. Lehmann, Measuring large-scale social networks with high resolution, PLoS ONE **9**, e95978 (2014).

[5] L. Apeltsin, J. H. Morris, P. C. Babbitt, and T. E. Ferrin, Improving the quality of protein similarity network clustering algorithms using the network edge weight distribution, Bioinformatics **27**, 326 (2011).

[6] E. Bullmore and O. Sporns, Complex brain networks: Graph theoretical analysis of structural and functional systems, Nat. Rev. Neurosci. **10**, 186 (2009).

[7] M. Rubinov and O. Sporns, Complex network measures of brain connectivity: Uses and interpretations, Neuroimage **52**, 1059 (2010).

[8] V. Sekara and S. Lehmann, The strength of friendship ties in proximity sensor data, PLoS ONE **9**, e100915 (2014).

[9] N. Langer, A. Pedroni, and L. Jäncke, The problem of thresholding in small-world network analysis, PLoS ONE **8**, e53199 (2013).

[10] W.-Q. Huang, X.-T. Zhuang, and S. Yao, A network analysis of the chinese stock market, Physica A **388**, 2956 (2009).

[11] M. Á. Serrano, M. Boguná, and A. Vespignani, Extracting the multiscale backbone of complex weighted networks, Proc. Natl. Acad. Sci. (USA) **106**, 6483 (2009).

[12] F. Radicchi, J. J. Ramasco, and S. Fortunato, Information filtering in complex weighted networks, Phys. Rev. E **83**, 046101 (2011).

[13] N. Dianati, Unwinding the hairball graph: Pruning algorithms for weighted complex networks, Phys. Rev. E **93**, 012304 (2016).

[14] J. Vaynberg and J. Qin, Weak protein-protein interactions as probed by NMR spectroscopy, Trends Biotechnol. **24**, 22 (2006).

[15] To see why $\rho > \frac{1}{2}$ is problematic, consider the marginal distribution for four edges, say $X_{ij}, X_{jk}, X_{kl}, X_{il}$. A simple calculation shows that the covariance matrix has a negative eigenvalue for $\rho > \frac{1}{2}$. Similarly, since $\mathrm{Var}[\Sigma_j X_{ij}]$ must be greater than 0, ρ must be greater than $-1/(n-2)$ and so negative correlations can be vanishingly weak at most.

[16] B. F. Maier, ThredgeCorr—A Python package for sampling from the locally correlated edge weight model, https://github.com/benmaier/ThredgeCorr.

[17] P. Diaconis and S. Janson, Graph limits and exchangeable random graphs, Rendiconti di Matematica, Serie VII **28**, 33 (2008).

[18] P. Orbanz and D. M. Roy, Bayesian models of graphs, arrays and other exchangeable random structures, IEEE Trans. Pattern Anal. Machine Intell. **37**, 437 (2015).

[19] D. N. Hoover, *Relations on Probability Spaces and Arrays of Random Variables* (Institute for Advanced Study, Princeton, NJ, 1979).

[20] B. Harris and A. P. Soms, The use of the tetrachoric series for evaluating multivariate normal probabilities, J. Multivar. Anal. **10**, 252 (1980).

[21] P. D. Miller, *Applied Asymptotic Analysis* (American Mathematical Society, 2006).

[22] K. M. Harris and J. R. Udry, National longitudinal study of adolescent health (Add Health), 1994-2008 [public use]. Carolina Population Center, University of North Carolina-Chapel Hill [distributor], Inter-university Consortium for Political and Social Research [distributor], 2018-08-06, https://doi.org/10.3886/ICPSR21600.v21.

[23] M. E. J. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. (USA) **98**, 404 (2001).

[24] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein, Reactome: A knowledge base of biological pathways, Nucl. Acids Res. **33**, D428 (2005).

[25] P. Erdős and A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hungarian Acad. Sci. **5**, 17 (1960).

[26] B. Bollobás, *Random Graphs*, 2nd ed., Cambridge Studies in Advanced Mathematics (Cambridge University Press, Cambridge, 2001).

[27] D. J. Watts and S. H. Strogatz, Collective dynamics of "small-world" networks, Nature (London) **393**, 440 (1998).

[28] D. Fernholz and V. Ramachandran, The diameter of sparse random graphs, Random Struct. Algorithms **31**, 482 (2007).

[29] R. Cohen and S. Havlin, Scale-Free Networks are Ultrasmall, Phys. Rev. Lett. **90**, 058701 (2003).

[30] R. Cohen, S. Havlin, and D. ben Avraham, Structural properties of scale-free networks, in *Handbook of Graphs and Networks*, edited by S. Bornholdt and H. G. Schuster (FRG: Wiley-VCH Verlag, Weinheim, 2004), pp. 85–110.

[31] Y. L. Tong, *The Multivariate Normal Distribution* (Springer-Verlag, New York, 1990).

[32] *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, edited by M. Abramowitz and I. A. Stegun (National Bureau of Standards, Washington, DC, 1964).