Sharpest possible clustering bounds using robust random graph analysis

Judith Brugman, Johan S.H. van Leeuwaarden and Clara Stegehuis

Complex network theory crucially depends on the assumptions made about the degree distribution, while fitting degree distributions to network data is challenging, in particular for scale-free networks with power-law degrees. We present a robust assessment of complex networks that does not depend on the entire degree distribution, but only on its mean, range and dispersion: summary statistics that are easy to obtain for most real-world networks. By solving several semi-infinite linear programs, we obtain tight (the sharpest possible) bounds for correlation and clustering measures, for all networks with degree distributions that share the same summary statistics. We identify various *extremal random graphs* that attain these tight bounds as the graphs with specific three-point degree distributions. We leverage the tight bounds to obtain robust laws that explain how degree-degree correlations and local clustering evolve as function of node degrees and network size. These robust laws indicate that power-law networks with diverging variance are among the most extreme networks in terms of correlation and clustering, building further theoretical foundation for widely reported scale-free network phenomena such as correlation and clustering decay.

I. INTRODUCTION

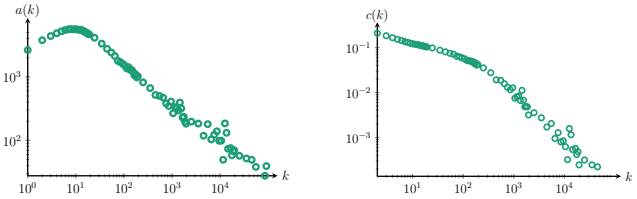
Degree heterogeneity drives many complex network properties, with the spread of a virus over a network as a striking example. In homogeneous networks, when differences between node connectivity are relatively small, classical network theory says an epidemic can arise when the average number of secondary infections caused by a single infected individual, R, exceeds one. In scale-free networks with high degree fluctuations, on the other hand, this is not a good predictor, as individuals who are infected early on may be different from the average individual. Indeed, these individuals typically have more contacts so that an epidemic can develop even if R is close to zero. A virus then spreads extremely quickly and can hardly be contained. Many real-world networks, in fact, often have extremely heterogeneous degrees that can be approximated with power-laws, so that the proportion of nodes having k neighbors scales as $k^{-\tau}$ with exponent τ between 2 and 3 [17, 26, 45]. Power-law degrees imply various intriguing scale-free network properties such as the absence of an epidemic threshold for $\tau < 3$ [24, 34], ultra-small distances [32] and efficient embedding methods [3].

Because of this degree heterogeneity, the analysis of such networks is complex. Network properties such as the friendship paradox, and more generally the connections between nodes with vastly different degrees, are studied in network theory in the form of so-called degree-degree correlations and clustering. Degree-degree correlations measure correlation between the degrees of two connected nodes, often captured in terms of $a(k)$, the average degree of a neighbor of a degree-k node. By clustering we mean the creation of triangular connections (triadic closure), quantified in terms of $c(k)$, the probability that two neighbors of a degree-k node are neighbors themselves. In uncorrelated networks the $a(k)$ and $c(k)$ are independent of k. However, the majority of real-world networks, and scale-free networks in particular, have $a(k)$ and $c(k)$ functions that decay in k, first observed in technological networks such as the Internet [33, 38]. Figure 1 shows the same fall-off for a social network: YouTube users as vertices, and edges indicating friendships between them [29].

When $a(k)$ decreases in k, the network is said to be disassortative, so that high-degree vertices typically connect to low-degree vertices. When $c(k)$ decreases in k, this may indicate the presence of hierarchy. A hierarchical topology arises, for example, when the rare high-degree nodes together form a backbone, and the low-degree nodes are located in clusters of low-degree nodes that are connected to one of the high-degree nodes. These core peripheries are found in complex networks created by both humans and nature [18]. This view of a hierarchical network explains both the negative degree-degree correlations, because most low-degree nodes are connected to a single high-degree node, and the clustering fall-off, because the core periphery mainly consists of triadic closures between low-degree communities while high-degree nodes rarely participate in triangles and communities.

Network features such as decaying degree correlations are broadly studied through random graphs, mathematically tractable models that can generate random samples of a graph in which nodes have i.i.d. degrees [2, 20, 21, 23, 42]. Random graph models take the degree distribution as input. Conditional on the degree distribution, random graph properties such as average distance and clustering can be characterized and tested against measurements from real-world network data with the same degree distribution.

Motivated by the wide range of examples of networks with heavy-tailed degrees, the power-law distribution has become a popular choice as an input degree distribution for random graph models. Fitting a power law to real-world data, however, is statistically challenging [8, 13, 46]. For small values, a power law is usually not a good fit. For this reason, lower bounds for the power-laws or additional slowly-varying functions are often introduced, but these form extra functions that need



(a) (b)

FIG. 1. a) $a(k)$ and b) $c(k)$ for the YouTube social network

to be fitted as well. Larger values of the power law also present challenges. Most real-world data sets only follow a power law up to some maximal degree, which is often modeled by an exponential cutoff [15, 30, 31]. Real-world networks are finite by definition, while a power law allows infinitely large values.

An inherent disadvantage of network theory that rests on random graphs is the dependency on precise statistical assumptions about the degree distributions. Network theory should not be overly sensitive to the assumed degree distribution, especially when the assumption is hard to justify statistically. For power laws for instance, the tail exponent τ implies vastly different network properties. One reason for this is the variance of the degree distribution. When the number of nodes n becomes large, the variance grows to infinity for $\tau < 3$, while the variance remains finite for $\tau > 3$. This difference in variance growth crucially influences the network structure and its degree-degree correlations [41, 47].

To overcome this sensitivity of network null models to precise statistical assumptions on the presence of power-laws or other specific degree distributions, we here characterize degree correlations and clustering in random graphs that only require partial information about the degree distribution. Inspired by the complicated assessment of power laws, we assume only the mean, dispersion and cutoff of the degree distribution are known. Here we consider two measures of dispersion: the variance and the mean absolute deviation (MAD) of the degree distribution. The MAD is an alternative to variance for measuring dispersion around the mean, and may be more appropriate in case of heavy tails. Indeed, MAD can deal with distributions that do not possess a finite variance, in particular the class of power-law distribution with $\tau \in (2, 3)$, for which MAD remains finite while variance becomes infinite in the large-network limit when $n \to \infty$.

We will establish the maximal correlation and maximal clustering that can be achieved by all degree distributions that share the same mean, cutoff and dispersion. By constructing and solving optimization problems, we find the extremal degree distributions that maximize the degree-degree correlation and clustering. These optimization problems take the form

$$\max_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\text{graph property}\right], \tag{1}$$

where $\mathcal{P}$ is the set that contains all degree distributions that comply with the limited information, such as the mean, cutoff and dispersion. Hence, within the set $\mathcal{P}$ we find the degree distribution $\mathbb{P}^*$ that maximizes the expected graph property. We will refer to the random graph with the extremal degree distribution that attains the maximum as the *extremal random graph*.

We solve optimization problems as in (1) for the hidden-variable model [6, 12], a random graph model that generates graphs with degrees that approximately follow some given distribution. The optimization problems in this paper give rise to semi-infinite linear programs and can be solved using methods from distributionally robust optimization. Using a primal-dual approach we can solve these semi-infinite linear programs in closed form, and find the precise description of the degree distribution $\mathbb{P}^*$ that attains the largest expected graph property. Since this distribution is by definition contained in the set $\mathcal{P}$, the bound

$$\mathbb{E}_{\mathbb{P}}\left[\text{graph property}\right] \leq \mathbb{E}_{\mathbb{P}^*}\left[\text{graph property}\right], \quad \forall \mathbb{P} \in \mathcal{P}. \tag{2}$$

is the best possible (tight) bound for all degree distributions that share the same summary statistics as is $\mathcal{P}$.

Distributionally robust optimization finds applications in many domains [1, 16, 36], but applications in the area of network science are rare. In fact, we are only aware of two papers that apply distributionally robust optimization to study complex networks. The first paper investigates the maximal possible subgraph counts under a restrictive cutoff scheme that creates uncorrelated networks [44]. The second paper provides a distributionally robust model for the influence maximization problem where the influence diffusion is adversarially adapted to the choice of seed set [10]. Here the authors aim to detect a seed set whose worst-case expected influence is maximized, and show that this differs from the standard model in which influence is assumed to diffuse independently across the different edges.

We term the largely unexplored approach taken in this paper *distributionally robust random graph analysis*, referring to the combination of classical random graph analysis and the optimization framework that only conditions on partial distributional information. This view on random graphs trades precise results that hold for a given degree distribution for robust statements that hold for classes of degree distributions. Such robust results fit well with the search for universal properties of complex networks.

Here are the main contributions of this paper:

(i) For all degree distributions with a given mean, variance and cutoff, we obtain the maximal degree-degree correlations and local clustering. We show that these bounds for $a(k)$ and $c(k)$ decay in k, as observed in most real-world networks and random graph models.

(ii) We show that extremal properties are often attained by uncorrelated graphs. As long as it is possible to create uncorrelated random graphs with the

given statistics, the most extreme random graph properties will be formed by uncorrelated graphs for all properties we investigate.

(iii) We compare the extremal graph models that provide the highest correlations and clustering to existing results for power-law random graphs. While power-laws are often thought of as degree distributions that lead to extreme behavior, the power-laws are not the degree distributions that possess the largest possible values of $a(k)$ and $c(k)$ when $\tau > 2$.

(iv) We provide a method to detect whether any given real-world data set can be modeled by hidden-variable models for properties of interest. We show that for several real-world data sets, no possible hidden-variable model can model the particular real-world data set.

We introduce the hidden-variable model and assumptions on the degree distribution in Section II. We then solve the maximization problem that finds the extremal random graph that generates the maximal degree-degree correlation in Section III. The scaling laws for clustering as function of the network size are presented in Section IV, and in Section V we do this for clique counts. In Section VI we obtain results for the setting when the MAD instead of the variance is used as dispersion measure. In Section VII we compare the robust bounds obtained in earlier sections with existing results for scale-free networks with power-law degrees, and with data from real-world networks.

II. RANDOM GRAPHS AND HIDDEN (RANDOM) VARIABLES

Our analysis of degree-degree correlations and clustering will be based on the hidden-variable model, a random graph model in which every vertex $i \in [n]$ has a weight h_i, and edges are formed between pairs of nodes with a probability that depends on both weights. More specifically, every pair of vertices is connected independently with probability

$$p(h_i, h_j) = \min\Big(\frac{h_i h_j}{h_s^2}, 1\Big) = \min\Big(\frac{h_i h_j}{\mu n}, 1\Big), \qquad (3)$$

where μ denotes the average weight, and h_s is the structural cutoff set to $\sqrt{\mu n}$ throughout this paper, in line with its typical choice for power-law networks [4, 6, 9, 14]. This choice of cutoff ensures that the weight of a vertex is close to its degree [4, 43]. The structural cutoff describes the maximal degree of vertices that are not prone to degree-degree correlations [6]. As soon as the degree of a vertex becomes larger than the structural cutoff, it is forced to connect to lower degree vertices, as only few high degree vertices can be present while keeping the average degree fixed.

The natural cutoff describes the constraint on the largest possible network degree, or the largest possible weight, h_c. In many real-world networks as well as in networks generated from power-law degrees, the largest observed degree is much larger than the structural cutoff of $\sqrt{\mu n}$. For example, when the degrees of the vertices follow a power-law distribution with degree-exponent τ, the largest degree scales as $n^{1/(\tau-1)}$. This means that the network contains vertices that are prone to degree-degree correlations and connection probabilities become non-convex, which makes the network analysis technically more challenging.

The hidden-variable model has several properties that make it amenable to analytical analysis. First of all, when the connection probabilities are chosen suitably, the weight h and the degree k of a vertex are similar with high probability. Indeed, the expected degree d_j of vertex j given its weight, satisfies

$$\begin{aligned}\mathbb{E}\left[d_j \mid h_j\right] &= \sum_{j\neq i} \min\Big(\frac{h_i h_j}{\mu n}, 1\Big) \\ &\approx \sum_{i\neq j} \frac{h_i h_j}{\mu n} \approx h_j. \end{aligned} \qquad (4)$$

To be more precise, when $h \gg 1$, then $k = h(1 + o(1))$ [20]. This makes it possible to interchange weights and degrees, which is convenient as the connection probabilities are defined in terms of weights. Secondly, when all hidden variables are assigned, most network statistics of interest can be computed as a function of the hidden variables. For example, the average degree of all neighbors of a vertex with weight h can be written as

$$a(h) = \frac{1}{h}\sum_{i=1}^{n} h_i \min\Big(\frac{h h_i}{\mu n}, 1\Big), \qquad (5)$$

where the sum is over all vertices in the network, and multiplies the weight of a vertex with the probability that vertex i connects to the weight-h vertex.

The local clustering coefficient denotes the probability that two randomly chosen neighbors of a vertex with weight h are neighbors themselves. This statistic can again be written as a function of the hidden variables. Formally,

$$c(h) = \frac{1}{h^2}\sum_{1\leq i<j\leq n} \min\Big(\frac{h h_i}{\mu n}, 1\Big)\min\Big(\frac{h h_j}{\mu n}, 1\Big)\min\Big(\frac{h_i h_j}{\mu n}, 1\Big). \qquad (6)$$

Here the sum is over all pairs of vertices in the network, and the term inside the summation computes the probability that these vertices form a triangle with the weight-h vertex.

At first sight, degree correlations and clustering and seem unrelated, as the former is defined in terms of edges

and the latter in terms of triplets of nodes. Still, intuitively $a(k)$ and $c(k)$ are related in the case of hidden-variable models. Indeed, the average neighbor of a vertex of a weight k vertex is $a(k)$. The probability that two such vertices connect scales as $a(k)^2/n$, when $a(k)$ is sufficiently small. This probability can be interpreted as the probability that two 'average' neighbors of a weight-k vertex connect. It turns out that this intuitive reasoning provides the correct scaling of $c(k)$ in some cases. That is, $c(k) \sim a(k)^2/n$ [41].

When studying real-world data sets, we can only observe $\bar{c}(k)$ and $\bar{a}(k)$, the local clustering coefficient and average degree of the neigbors of a degree-k vertex, rather than a weight-h vertex. Still, the property of the hidden-variable model that degrees and weights are close makes the difference between these two statistics small in the large-network limit [20].

In traditional hidden-variable models, the weights $h_1, \dots, h_n$ are assumed independent and following some distribution $\mathbb{P}$. The natural cutoff can then be calculated from the distribution $\mathbb{P}$. In this paper, however, we only specify partial information about the weight (i.e. degree) distribution. We will assume that for the weights we know the minimal value, their maximal value (the natural cutoff h_s), the mean and the dispersion, first measured in variance and later in terms of mean absolute deviation (MAD). Let h denote a generic weight. We first assume that the weights are sampled independently from a distribution such that (i) $h = h_i$ has support $\mathrm{supp}(h) = [a, h_c]$ with $-\infty < a \leq h_c < \infty$, (ii) $\mathbb{E}[h] = \mu$ and (iii) $\mathbb{E}\left[(h-\mu)^2\right] = \sigma^2$. This defines the ambiguity set

$$\begin{aligned}\mathcal{P}(\mu,\sigma) =& \{\mathbb{P} : \mathrm{supp}(h) \subseteq [a, h_c], \mathbb{E}[h] = \mu, \\ & \mathbb{E}\left[(h-\mu)^2\right] = \sigma^2\}. \end{aligned} \tag{7}$$

Hence, when we now analyze the hidden-variable model under the assumption that the weight distribution belongs to $\mathcal{P}(\mu,\sigma)$, we perform a distributionally robust analysis of the random graph model.

The variance of the degree distribution is often highly affected by the choice of the natural cutoff. In power-law random graphs for example, the variance σ^2 grows as $h_c^{3-\tau}$. Indeed, for $\tau \in (2,3)$, the variance of the weights can be computed as

$$\int_1^{h_c} x^{2-\tau} dx - \mu^2 \sim h_c^{3-\tau}. \tag{8}$$

The MAD on the other hand, always satisfies the inequality $d < 2\mu$. Thus, as long as the average degree is finite, the MAD will not grow as a function of h_c.

In the rest of this paper, we will focus on the graph properties mentioned above, and first seek for the weight distribution $\mathbb{P}$ that solves

$$\max_{\mathbb{P}\in\mathcal{P}(\mu,\sigma)} \mathbb{E}_{\mathbb{P}}[\text{graph property}] \tag{9}$$

with $\mathcal{P}(\mu,\sigma)$ as in (7). This means that we take a distributionally robust approach for the input weights of the hidden-variable model under the assumption that their distribution belongs to $\mathcal{P}(\mu,\sigma)$. When the graph property (9) can be viewed as a convex function of the generic weight random variable h, (9) is optimized for a specific distribution with support on three points [44]. Indeed, due to the convex nature of the function, an optimizer aims to put as much weights on the extremal points a and h_c, while still adhering to the constraints on the average weight and its variance. This leads to a specific three-point distribution with probability mass on a, h_c and μ. However, in the setting we now consider with natural cutoff $h_c > \sqrt{\mu n}$, the connection probability (3) is not convex, and therefore most graph properties will also not be convex in the hidden variables. In the next sections, we therefore apply a primal-dual based approach to find the distributionally robust graph properties.

III. ROBUST DEGREE-DEGREE CORRELATIONS BOUNDS

The definition of $a(h)$ in (5) assumes that the hidden variables are known. Instead, we now assume that all hidden variables are drawn from some probability distribution $\mathbb{P}$, so that the expected value of $a(h)$ can be computed as

$$\mathbb{E}_{\mathbb{P}}[a(h)] = \frac{n}{h}\mathbb{E}_{\mathbb{P}}\left[h' \min\left(\frac{hh'}{\mu n}, 1\right)\right]. \tag{10}$$

We then search for the weight distribution $\mathbb{P}$ that solves

$$\max_{\mathbb{P}\in\mathcal{P}(\mu,\sigma)} \mathbb{E}_{\mathbb{P}}[a(h)] \tag{11}$$

with $\mathcal{P}(\mu,\sigma)$ as in (7). Hence, when we now analyze the hidden-variable model under the assumption that the weight distribution belongs to $\mathcal{P}(\mu,\sigma)$, we perform a robust analysis for all distributions with a given mean, variance and cutoff. The optimization problem (11) can be written as

$$\begin{aligned} \max_{\mathbb{P}(x)\geq 0} \quad & \int_x g(x)\mathrm{d}\mathbb{P}(x) \\ \text{s.t.} \quad & \int_x x^2 \mathrm{d}\mathbb{P}(x) = \mu^2 + \sigma^2, \ \int_x x \mathrm{d}\mathbb{P}(x) = \mu, \\ & \int_x \mathrm{d}\mathbb{P}(x) = 1, \end{aligned} \tag{12}$$

where $g(x) = x\min(hx/(\mu n), 1)$. In optimization theory, (12) is called a semi-infinite linear optimization problem (LP). The Richter-Rogosinski Theorem (see, e.g., [19, 39, 40]) says there exists an extremal distribution for problem (12) with at most three support points. While finding these points in closed form is typically not possible for general semi-infinite problems, we next show that this is possible for the problem at hand by resorting to the dual problem; see e.g. [22] and [36]. This dual

problem of (12) is given by

$$\begin{aligned} \min_{\lambda_1,\lambda_2,\lambda_3} \quad & \lambda_1(\mu^2+\sigma^2)+\lambda_2\mu+\lambda_3 \\ \text{s.t.} \quad & g(x)-\lambda_1 x^2-\lambda_2 x-\lambda_3 \leq 0 \quad \forall x\in[a,h_c], \end{aligned} \tag{13}$$

and aims to find a tightest quadratic majorant of $g(x)$ that minimizes $\lambda_1(\mu^2+\sigma^2)+\lambda_2\mu+\lambda_3$. Now $g(x)$ has a quadratic part up to $\min(\mu n/h, h_c)$, and a linear part. Figure 2 shows that this function has two possible tightest quadratic majorants. The first, $F_1(x)$, is given by $\lambda_1 = h/(\mu n), \lambda_2 = \lambda_3 = 0$ and has objective value $h(\mu^2+\sigma^2)/(\mu n)$. The second one, $F_2(x)$, is given by $\lambda_2 = 1, \lambda_1 = \lambda_3 = 0$, and has objective value μ. Which of the two majorants has the smallest objective value depends on h. For low values of h, the first majorant gives the lowest objective value, whereas for high values of h, the linear one dominates. The changing point is when

$$h = \mu^2 n/(\mu^2+\sigma^2). \tag{14}$$

The next step is to find a feasible solution for the primal problem that yields the same objective value as the solution to the dual problem. By weak duality of semi-infinite linear programming, a feasible solution to the dual problem gives a valid upper bound for the optimal primal solution value. A feasible primal solution with an objective value equal to this upper bound would show that strong duality holds. Next, we provide a constructive approach, based on the condition of complementary slackness, to find such a primal solution.

Assume that strong duality holds. The primal and dual objective are then equal for the primal maximizer $\mathbb{P}^*$ and the dual minimizer $(\lambda_1^*, \lambda_2^*, \lambda_3^*)$, and we can substitute the primal constraints in the dual objective. Hence, we obtain the relation

$$\int_x g(x)\mathrm{d}\mathbb{P}^*(x) = \int_x \lambda_1^* x^2+\lambda_2^* x+\lambda_3^* \,\mathrm{d}\mathbb{P}^*(x). \tag{15}$$

Since $\lambda_1(\mu^2+\sigma^2)+\lambda_2\mu+\lambda_3 - g(x) \geq 0$ pointwise by weak duality, (15) implies that $\mathbb{P}^*$ is supported only on the points where $\lambda_1^* x^2+\lambda_2^* x+\lambda_3^*$ coincides with $g(x)$. We now show that in both cases, a three-point distribution achieves the dual objective value.

In the first case, $h < \mu^2 n/(\mu^2+\sigma^2)$, and the dual objective value is given by $h(\mu^2+\sigma^2)/(\mu n)$. Now consider the three-point distribution (for ease of notation, we assume that $a = 0$, and denote $l = \min(\mu n/h, h_c)$) on the points $0, \mu, l$, so indeed the quadratic majorant and $g(x)$ coincide. The probabilities are chosen such that the moment conditions are satisfied. We obtain

$$p_0 = \frac{\sigma^2}{l\mu}, \quad p_\mu = 1-\frac{\sigma^2}{l\mu}-\frac{\sigma^2}{l(l-\mu)}, \quad p_l = \frac{\sigma^2}{l(l-\mu)}. \tag{16}$$

This is only a proper probability distribution when $\sigma^2/(l\mu)+\sigma^2/(l(l-\mu)) \leq 1$, which can be rewritten as $\sigma^2+\mu^2 \leq \mu l$. We first assume that $\min(\mu n/h, h_c) = \mu n/h$, so we need to check that $\sigma^2+\mu^2 \leq \mu^2 n/h$.

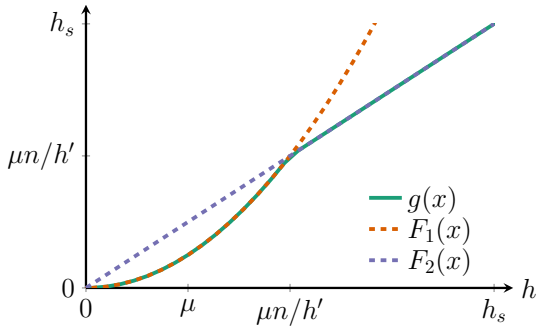


FIG. 2. The two possible tightest majorants for the function $g(x)$

This follows directly from the assumption on h. When $\min(\mu n/h, h_c) = h_c$, we should satisfy $\sigma^2+\mu^2 \leq \mu h_c$. This is always the case, as the maximal variance of a primal solution with mean μ is given by the primal solution $1-p_0 = p_{h_c} = \mu/h_c$, giving as variance $\sigma^2 \leq \mu h_c - \mu^2$.

This three-point distribution gives as objective value for the primal problem

$$\left(1-\frac{\sigma^2}{l\mu}-\frac{\sigma^2}{l(l-\mu)}\right)\frac{\mu^2 h}{\mu n}+\frac{\sigma^2}{l(l-\mu)}\frac{l^2 h}{\mu n} = \frac{(\sigma^2+\mu^2)h}{\mu n}, \tag{17}$$

which is equal to the dual objective value.

Thus, by duality, this three-point degree distribution generates the extremal random graph for $a(k)$, and the result that for $h < \mu^2 n/(\mu^2+\sigma^2)$,

$$\max_{\mathbb{P}\in\mathcal{P}^*(\mu,\sigma^2)} \mathbb{E}_{\mathbb{P}}[a(h)] = \frac{\mu^2+\sigma^2}{\mu}. \tag{18}$$

In the second case, $h > \mu^2 n/(\mu^2+\sigma^2)$, and the dual objective value is given by μ. Here, consider the three-point distribution

$$\begin{aligned} p_0 &= 1-p_{\mu n/h}-p_{h_c}, \quad p_{\mu n/h} = \frac{h^2\left(-h_c\mu+\mu^2+\sigma^2\right)}{\mu n(\mu n-h_c h)}, \\ p_{h_c} &= \frac{\mu^2(h-n)+h\sigma^2}{h_c(h_c h-\mu n)}. \end{aligned} \tag{19}$$

This is only a proper distribution when $\sigma^2 < h_c\mu-\mu^2$, which always holds by a similar reasoning as for the first case. The second condition of $\sigma^2 > h_c\mu-\mu^2-\mu n/h^2(h_c h-\mu n)$ is more involved, and is in fact not necessary. Indeed, in (19), we chose the third point of the distribution at h_c. However, as the tightest majorant in Figure 2 touches on an entire line, it is also possible to choose the third point somewhere else in $[\mu n/h, h_c]$ while achieving the same optimal value. Choosing another point for the three-point distribution also results in different conditions on σ^2. The lowest such constraint is when $p_0 = 1-p_{\mu n/h}$ in (19), yielding $\sigma^2 = \mu^2 n/h-\mu^2$, which is ensured by our assumption on h. Under this three-point distribution, $\mathbb{E}[X] = \mu$, $\mathbb{E}[(X-\mu)^2] = \sigma^2$ and the primal objective value $\mathbb{E}[g(X)] = \mu$.

Thus, by duality, the three-point distribution is the variance-based extremal random graph for $a(k)$, giving that for $h > \mu^2 n/(\mu^2+\sigma^2)$,

$$\max_{\mathbb{P}\in\mathcal{P}(\mu,\sigma^2)} \mathbb{E}_{\mathbb{P}}\left[a(h)\right] = \frac{\mu n}{h}. \tag{20}$$

Notice that the three-point distribution (19) is not a unique optimum, as the dual function $F_2(x)$ coincides with $g(x)$ on the entire interval $[\mu n/h, h_c]$. Therefore, one can construct an arbitrary (discrete, continuous or mixed) probability distribution with support on the interval $[\mu n/h, h_c]$, as long as the mean and variance conditions are satisfied. Similarly, the three-point distribution (16) is also not unique.

This yields the following theorem:

Theorem III.1. *When* $p(h,h') = \min(hh'/(\mu n), 1)$*, and* $h_c \ll n$*,*

$$\max_{\mathbb{P}\in\mathcal{P}(\mu,\sigma^2)} \mathbb{E}_{\mathbb{P}}\left[a(h)\right] = \begin{cases} \frac{\mu^2+\sigma^2}{\mu} & h < \mu^2 n/(\mu^2+\sigma^2) \\ \frac{\mu n}{h} & h > \mu^2 n/(\mu^2+\sigma^2). \end{cases} \tag{21}$$

The theorem distinguishes two regimes: one constant regime for low h, and a decaying regime for high enough h. In the constant regime, vertices are not prone to degree-degree correlations: all vertices have the same average nearest neighbor degree as long as $h < \mu^2 n/(\mu^2+\sigma^2)$. Furthermore, higher degree variance implies a lower threshold, and hence more vertices that are subject to degree-degree correlations. This is consistent with the intuition that degree-degree correlations arise because high-degree vertices are forced to connect to lower-degree vertices due to the lack of sufficiently many high-degree vertices. When $h_c < \mu^2 n/(\mu^2+\sigma^2)$, it is possible to create entirely uncorrelated networks. This condition is more general than the often used $h_c \ll \sqrt{n}$ constraint for uncorrelated networks that was found for power-law networks [6], as here we do not assume any specific weight distribution.

IV. ROBUST CLUSTERING BOUNDS

We now consider the probability that two randomly chosen neighbors of a vertex of weight h are connected to one another as well, $c(h)$. Again, in (6) the hidden variables are assumed to be fixed. We assume that the hidden variables are drawn from some distribution $\mathbb{P}$, so that the expected value of $c(h)$ becomes

$$\mathbb{E}_{\mathbb{P}}\left[c(h)\right] = \frac{n^2}{h^2}\mathbb{E}_{\mathbb{P}}\left[\min\left(\frac{hh_1}{\mu n},1\right)\min\left(\frac{hh_2}{\mu n},1\right)\min\left(\frac{h_1h_2}{\mu n},1\right)\right] \tag{22}$$

Now, instead of optimizing over only one hidden-variable as in (12), we need to jointly optimize the weight distributions of both neighbors. Under variance-based optimization, the optimization problem for $c(h)$ corresponding to (12) becomes

$$\begin{aligned} \max_{\mathbb{P}(x)\geq 0} \quad & \int_{x_1}\int_{x_2} g(x_1,x_2)\mathrm{d}\mathbb{P}(x_2)\mathrm{d}\mathbb{P}(x_1) \\ \text{s.t.} \quad & \int_x x^2\mathrm{d}\mathbb{P}(x) = \mu^2+\sigma^2, \int_x x\mathrm{d}\mathbb{P}(x) = \mu, \\ & \int_x \mathrm{d}\mathbb{P}(x) = 1, \end{aligned} \tag{23}$$

where

$$g(x_1,x_2) = \min\left(\frac{x_1x_2}{\mu n},1\right)\min\left(\frac{x_1h}{\mu n},1\right)\min\left(\frac{x_2h}{\mu n},1\right). \tag{24}$$

It turns out that this optimization problem is difficult to solve, due to the constraint that the two variables x_1 and x_2 are i.i.d.. We therefore instead solve a relaxation of (23), where we allow these two variables to be correlated, or drawn from different distributions. This relaxed problem takes the form

$$\begin{aligned} \max_{\mathbb{P}(x_1,x_2)\geq 0} \quad & \int_{x_1}\int_{x_2} g(x_1,x_2)\mathrm{d}\mathbb{P}(x_1,x_2) \\ \text{s.t.} \quad & \int_{x_1}\int_{x_2} x_1^2x_2^2\mathrm{d}\mathbb{P}(x_1,x_2) = (\mu^2+\sigma^2)^2, \\ & \int_{x_1}\int_{x_2} x_1x_2\mathrm{d}\mathbb{P}(x_1,x_2) = \mu^2, \\ & \int_{x_1}\int_{x_2} \mathrm{d}\mathbb{P}(x_1,x_2) = 1. \end{aligned}$$

Here, instead of optimizing over a single distribution from which both weights are drawn, we optimize over a joint, symmetric distribution over the weights of the other two vertices involved in a triangle, $\mathbb{P}(x_1,x_2)$. As a consequence, it is possible that the optimal weight distributions found by solving the relaxed problem are correlated so that the two vertices jointly optimize their weights to make a triangle formation more likely. Interestingly, it turns out that this is not the case. In Appendix A 1, we show that the optimizer of the relaxed optimization problem, that thus allows for correlations and is easier to solve, is in fact an uncorrelated distribution, and therefore also the solution of the original optimization problem (23). We are able to derive this optimizer, because the dual version of this problem

$$\begin{aligned} \min_{\lambda_1,\lambda_2,\lambda_3} \quad & \lambda_1(\mu^2+\sigma^2)^2 + \lambda_2\mu^2 + \lambda_3 \\ \text{s.t.} \quad & g(x_1,x_2) - \lambda_1x_1^2x_2^2 - \lambda_2x_1x_2 - \lambda_3 \leq 0 \\ & \forall x_1,x_2 \in [a,h_c], \end{aligned}$$

is relatively easy to solve. This gives the following theorem on the distributionally robust optimizer for $c(h)$:

Theorem IV.1. *When* $p(h,h') = \min(hh'/(\mu n), 1)$ *and* $\sigma^2 < \mu \max(\sqrt{\mu n}, \mu n/h) - \mu^2$,

$$\max_{\mathbb{P}\in\mathcal{P}(\mu,\sigma^2)} \mathbb{E}_{\mathbb{P}}\left[c(h)\right] = \begin{cases} \min\left(\frac{(\mu^2+\sigma^2)^2}{\mu^3 n}, 1\right) & h < \mu^2 n/(\mu^2+\sigma^2) \\ \frac{\mu n}{h^2} & h > \mu^2 n/(\mu^2+\sigma^2). \end{cases} \quad (25)$$

This theorem only holds when σ^2 is not too large. We conjecture that when σ^2 is larger than the range prescribed by the theorem, a primal-dual gap appears, indicating that the optimization problem cannot be solved anymore through primal-dual methods. Indeed, for larger σ^2, the best dual solution remains feasible as it does not depend on σ^2. However, it is then impossible to construct a probability distribution with the required variance on the set of values where the dual constraints are tight. This suggests that a primal-dual gap is present in that case, as there is no primal feasible solution that satisfies complementary slackness. For large σ^2, this implies that the primal problem has to be solved without the help of the dual problem, which makes the problem significantly more challenging.

V. ROBUST CLIQUE COUNTS

Whereas $a(k)$ and $c(k)$ measure two- and three-point correlations between nodes, we demonstrate in this section that robust bounds can also be obtained for network statistics that include more than three nodes. We focus on the number of cliques of size k, denoted as $N(K_k)$, and use that

$$\begin{aligned}\mathbb{E}\left[N(K_k)\right] &= \sum_{1\le i_1<i_2\cdots<i_k} \prod_{u<v} p(h_{i_u}, h_{i_v}) \\ &= \sum_{1\le i_1<i_2\cdots<i_k} \prod_{u<v} \min\Big(\frac{h_{i_u}h_{i_v}}{\mu n}, 1\Big), \end{aligned} \quad (26)$$

as a clique is present if and only if all possible edges between nodes $i_1, i_2 \ldots, i_k$ are present. To establish the variance-based bound on the expected number of cliques, we formulate the multivariate optimization problem

$$\begin{aligned} \max_{\mathbb{P}(x)\ge 0} \quad & \int_{x_1}\cdots\int_{x_k} g(x_1,x_2,\ldots,x_k)\mathrm{d}\mathbb{P}(x_1)\ldots\mathrm{d}\mathbb{P}(x_k) \\ \text{s.t.} \quad & \int_x x^2 \mathrm{d}\mathbb{P}(x) = \mu^2+\sigma^2, \int_x x\mathrm{d}\mathbb{P}(x) = \mu, \\ & \int_x \mathrm{d}\mathbb{P}(x) = 1 \end{aligned} \quad (27)$$

with

$$g(x_1,\ldots,x_k) = \prod_{1\le i<j\le k} \min\Big(\frac{x_i x_j}{\mu n}, 1\Big).$$

As for clustering, this optimization problem appears intractable due to the i.i.d. hidden variables. We therefore again solve a relaxation of (27) instead that drops the i.i.d. assumption. In Appendix A 1 we solve this relaxed optimization problem and prove the following robust bounds for clique counts:

Theorem V.1. *When* $\sigma^2 \le \mu(\sqrt{\mu n} - \mu)$, $k > 3$, *and as* $n \to \infty$,

$$\max_{\mathbb{P}\in\mathcal{P}(\mu,\sigma^2)} \mathbb{E}_{\mathbb{P}}\left[N(K_k)\right] = \frac{(\mu^2+\sigma^2)^k}{\mu^k k!}(1+o(1)). \quad (28)$$

Furthermore, for $k = 3$, $\sigma^2 \le \mu(\sqrt{\mu n} - \mu)$ *and for any* n,

$$\max_{\mathbb{P}\in\mathcal{P}(\mu,\sigma^2)} \mathbb{E}_{\mathbb{P}}\left[N(K_k)\right] = \frac{(\mu^2+\sigma^2)^k}{\mu^k k!}. \quad (29)$$

This theorem shows that cliques significantly increase when the variance grows, as is the case for heavy-tailed weight distributions. The theorem only holds asymptotically (except for $k = 3$), as we create primal and dual solutions with a small gap between their respective optimal values that vanishes as $n \to \infty$. Whereas Theorems III.1and IV.1 gave exact (non-asymptotic) results, here the relaxed optimization method that provided exact results for Theorem IV.1 gives non i.i.d. weight distributions. Thus, this method does not provide exact bounds on clique counts. Instead, we first solve the dual problem, and then construct i.i.d. primal weight distributions that asymptotically achieve the dual value, and are therefore asymptotically optimal.

As in Theorem IV.1, the theorem contains a condition on σ^2. We conjecture that for larger σ^2, a larger primal-dual gap is present that does not vanish in the large-network limit, so that the optimization problem (27) cannot be solved through its dual variant, similarly as for Theorem IV.1.

VI. MAD INSTEAD OF VARIANCE

We now turn to a second measure of dispersion: mean absolute deviation. While the variance of a random variable can be infinite, the MAD is always bounded by 2μ, so that even in networks with heavy-tailed degree distributions this quantity remains finite. For maximizing based on MAD, the ambiguity set now becomes

$$\begin{aligned}&\mathcal{P}(\mu, d) = \\ &\{\mathbb{P} : \mathrm{supp}(h) \subseteq [a, h_c], \mathbb{E}\left[h\right] = \mu, \mathbb{E}\left[|h-\mu|\right] = d\}. \end{aligned} \quad (30)$$

As for the variance-based approach, we then aim to find the probability distribution $\mathbb{P} \in \mathcal{P}(\mu, d)$ that maximizes the network statistics $a(h)$ and $c(h)$. We can use the same approach of constructing an optimization problem based on the constraints formed by the ambiguity set and finding the optimal primal-dual solution. As shown in Appendix A 2, we obtain the following result for $a(h)$:

Theorem VI.1. *When* $p(h,h') = \min(hh'/(\mu n), 1)$, *and* $h_c \to \infty$ *as* $n \to \infty$ *and* $h_c \ll n$,

$$\max_{\mathbb{P}\in\mathcal{P}(\mu,d)} \mathbb{E}_{\mathbb{P}}\left[a(h)\right] = \frac{d}{2\mu}\min\left(h_c, \frac{\mu n}{h}\right)(1+o(1)). \quad (31)$$

This optimal value of $a(h)$ is attained by the 3-point distribution

$$p_0 = \frac{d}{2\mu}, \quad p_\mu = 1 - \frac{d}{2\mu} - \frac{d}{2(l-\mu)}, \quad p_l = \frac{d}{2(l-\mu)}, \quad (32)$$

where $l = \min(\mu n/h, h_c)$.

To obtain results for $c(k)$ with MAD as well, we need to solve the optimization problem

$$\begin{aligned} \max_{\mathbb{P}(x)\geq 0} \quad & \int_{x_1}\int_{x_2} g(x_1,x_2)\mathrm{d}\mathbb{P}(x_2)\mathrm{d}\mathbb{P}(x_1) \\ \text{s.t.} \quad & \int_x |x-\mu|\mathrm{d}\mathbb{P}(x) = d, \int_x x\mathrm{d}\mathbb{P}(x) = \mu, \\ & \int_x \mathrm{d}\mathbb{P}(x) = 1, \end{aligned} \quad (33)$$

similarly to (23). Again, this optimization problem is difficult to solve, due to the constraint that the two variables x_1 and x_2 are i.i.d.. We could solve this problem by proceeding as in Section IV, by writing an unconstrained version of this optimization problem, where we allow the two variables to be non-identical or correlated. However, these techniques then lead to the dual problem

$$\begin{aligned} &\min_{\lambda_1,\lambda_2,\lambda_3} \lambda_1 d^2 + \lambda_2\mu^2 + \lambda_3 \\ &\text{s.t. } g(x_1,x_2) - \lambda_1|x_1-\mu||x_2-\mu| - \lambda_2 x_1 x_2 - \lambda_3 \leq 0 \\ &\forall x_1, x_2 \in [a, h_c]. \end{aligned}$$

Compared with (25), this dual function is no longer quadratic, but a product of absolute values summed with a linear term. The dual then tries to find the tightest majorant of this two-dimensional function of the non-convex function $g(x_1,x_2)$ illustrated in Figure 3. However, finding the tightest majorant of $\lambda_1|x_1-\mu||x_2-\mu| - \lambda_2 x_1 x_2 - \lambda_3$ to a general function is not obvious, as it is a function with a kink and different behavior near the x and y axes from the line $y = x$, as illustrated in Figure 4. This makes it more difficult to find the values of $\lambda_1, \lambda_2, \lambda_3$ that find this tightest majorant.

Therefore, we take a different approach instead. We first focus on the one-dimensional problem: for fixed x_1, what is the distribution of x_2 that maximizes $g(x_1,x_2)$? This problem can now be solved by a one-dimensional dual problem, which is easy to solve, similarly as in Section III. We then take the distribution of the optimal x_2 (which may depend on the value of x_1), and then optimize over the distribution of x_1 as well. Now this iterative approach may introduce correlations between the distributions: it is possible that the optimal distribution for x_1 is different from, or dependent on, the optimal distribution of x_2. Still, in some cases it may be true that

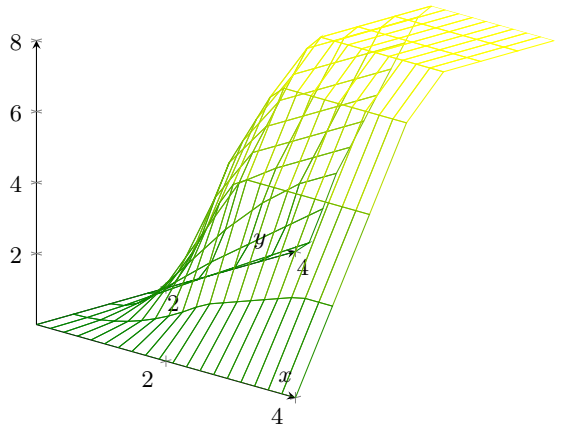


FIG. 3. The function $g(x,y)$.

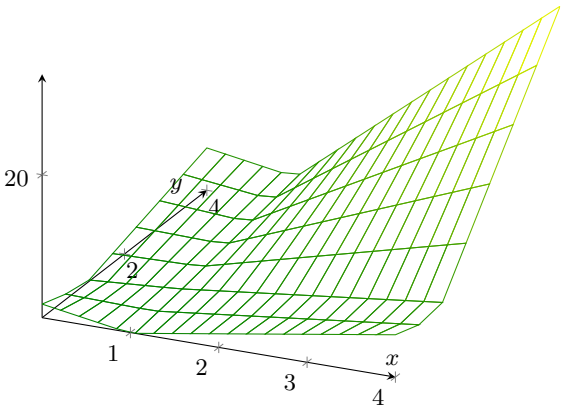


FIG. 4. The function $2|x-1||y-1| + xy$.

the output of this less restrictive optimization problem gives an i.i.d. distribution of x_1 and x_2. In that case, this method also solves (33). In the case when $g(x_1,x_2)$ is convex, this is known to be true. Unfortunately, for the case of computing $c(h)$, $g(x_1,x_2)$ is not convex. And indeed, the iterative optimizer is not always i.i.d. in this case. Still, we show that asymptotically, an i.i.d. optimizer achieves the same maximal value of $c(h)$. Applying this iterative method, as shown in Appendix A 2, leads to the following results:

Theorem VI.2. *When* $p(h,h') = \min(hh'/(\mu n), 1)$ *and* $h_c \to \infty$ *as* $n \to \infty$,

$$\max_{\mathbb{P}\in\mathcal{P}(\mu,d)} \mathbb{E}_{\mathbb{P}}\left[c(h)\right] = \begin{cases} 1 & h \ll \sqrt{\mu n} \text{ and} \\ & 2\mu(1-\sqrt{\mu/n}) < d < 2\mu(1-h/n) \\ \frac{d^2}{4\mu^2}(1+o(1)) & h \ll \sqrt{\mu n} \text{ and } d < 2\mu(1-\sqrt{\mu/n}) \\ \frac{d^2 n}{4\mu h^2}(1+o(1)) & \sqrt{\mu n} \ll h \ll n. \end{cases} \quad (34)$$

While from the perspective of i.i.d. weight sampling of the random graph it is natural to constrain the two

vertices that form a triangle together with the degree-h node to be sampled from the same distribution, in the optimization problem it is also possible to find the optimal pair of correlated distributions over the two nodes that form a triangle together with the weight-h node. In that case, the nodes that form a triangle together with the weight-h node are both sampled from the ambiguity set $\mathcal{P}_{(\mu,d)}$, so that they still both have mean weight μ and MAD d. However, the weights of the two nodes are now allowed to be correlated.

In the proof of Theorem VI.2, we show that in the ranges where Theorem VI.2 is valid, adding correlations between the distributions of h_1 and h_2 does not make a difference in the scaling for $c(h)$. Indeed, the upper bound of 1 is always valid for $c(h)$, so that lifting the constraint on the i.i.d. distributions of h_1 and h_2 cannot increase the optimal value of $c(h)$. For the second setting where $\sqrt{\mu n} \ll h \ll n$, we in fact show in the proof of Theorem VI.2 that by optimizing $c(h)$ over a set of probability distributions on h_1 and h_2 which are allowed to be correlated, we end up with i.i.d. distributions as the optimizer. So here also lifting the constraint on the joint distribution will not increase $c(h)$. Furthermore, in the case $h \ll \sqrt{\mu n}$ and d is small, allowing correlations only increases the non-leading order terms, so that asymptotically, $c(h)$ cannot be increased by allowing correlations between the distributions of h_1 and h_2.

VII. COMPARISON WITH OTHER NETWORKS

We now compare the extremal values of $a(k)$ and $c(k)$ to several synthetic and real network data.

a. Erdős-Rényi model. In the Erdős-Rényi model, $h_i = \mu, \forall i$, and all vertices connect with probability μ/n. Thus, $\sigma^2 = 0$ and $d = 0$. The zero variance and MAD means that there is only one distribution in the ambiguity sets (7) and (30), which is $h_i = \mu, \forall i$. Therefore, Theorem III.1 and VI.1 predict that the maximal value of $a(h) = \mu$, which is the exact average weight of a neighbor in an Erdős-Rényi random graph, as all vertices have weight μ. Furthermore, Theorem IV.1 and VI.2 predict that the maximal value of $c(h) = \mu/n$, which is also equal to $c(h)$ in an Erdős Rényi model, as the probability that two neighbors of a vertex connect is μ/n, the same as the connection probability for all pairs of vertices. Thus, for Erdős-Rényi random graphs, our bounds are tight.

b. Poisson random graph. When the hidden-variables have a Poisson distribution with mean μ, the second moment of the weight distribution is $\mu + \mu^2$. For such networks, $hh' < \mu n$ almost surely for all $h \ll n$. Thus, (10) gives $\mathbb{E}[a(h)] = 1 + \mu$. Applying Theorem III.1 yields that the maximal possible $a(h)$ value among all networks with the same mean and variance of the weights also equals $1 + \mu$. Similarly $\mathbb{E}[c(h)] = (\mu + \mu^2)^2/(\mu^3 n)$ for Poisson random graphs by Equation (22), while by Theorem IV.1, the maximal value of $c(k)$ among all power-law random graphs also equals $(\mu + \mu^2)^2/(\mu^3 n)$. Thus, Poisson random graphs achieve the maximal bounds for $a(h)$ and $c(h)$ exactly.

When looking at the MAD-based bounds, the picture changes drastically. Indeed, for a Poisson random variable with integer mean μ [37]

$$d = 2\mu^{\mu+1} e^{-\mu}/\mu!, \tag{35}$$

so that by Theorem VI.1 the maximal value of $a(h)$ scales for low h as $h_c \mu^\mu e^{-\mu}/\mu!$, which can be much larger than the achieved value of $1 + \mu$.

For $c(k)$, Theorem VI.2 yields that for h low, the maximal value of $c(h)$ for random graphs with the same mean and MAD as the Poisson random graph equals $(\mu^\mu e^{-\mu}/\mu!)^2$, which is an n-independent constant, in contrast to the achieved value of $c(h) = (\mu + \mu^2)^2/(\mu^3 n)$, which decays in n. Thus, our variance-based bounds can be significantly lower than the ones based on equal MAD.

c. Comparing power-law $a(h)$ to extremal $a(h)$. We now turn to random graphs with power-law distributed weights. We first compare the maximal scaling of $a(h)$ given by Theorem III.1 to the value of $a(h)$ attained by the power-law weight distribution

$$\mathbb{P}(h > x) = C x^{1-\tau}. \tag{36}$$

When sampling n i.i.d. weights from this distribution, the maximal weight scales as $h_c = n^{1/(\tau-1)}$ with high probability. In such power-law Chung-Lu models [41, 47],

$$a(h) \sim \begin{cases} n^{(3-\tau)/(\tau-1)} & h \ll n^{(\tau-2)/(\tau-1)} \\ (n/h)^{3-\tau} & h \gg n^{(\tau-2)/(\tau-1)}. \end{cases} \tag{37}$$

We now investigate how close this value of $a(h)$ is to the maximal possible values among all Chung-Lu models with the same mean and variance as the power-law distribution. For power-law distributed weights, $\sigma^2 \sim n^{(3-\tau)/(\tau-1)}$, as derived in (8). Thus, Theorem III.1 yields

$$\begin{aligned} \max_{\mathbb{P} \in \mathcal{P}(\mu, n^{(3-\tau)/(\tau-1)})} \mathbb{E}_{\mathbb{P}}[a(h)] &= \frac{\mu^2 + n^{(3-\tau)/(\tau-1)}}{\mu} \\ &\sim n^{(3-\tau)/(\tau-1)}, \end{aligned} \tag{38}$$

when $h < n^{(2\tau-4)/(\tau-1)}$, while

$$\max_{\mathbb{P} \in \mathcal{P}(\mu, n^{(3-\tau)/(\tau-1)})} \mathbb{E}_{\mathbb{P}}[a(h)] = \frac{\mu n}{h}, \tag{39}$$

when $h > n^{(2\tau-4)/(\tau-1)}$ For large h, the scaling in (37) only agrees with the value of $\mu n/h$ of (39) for $\tau = 2$. For low h, the power-law value of $a(h)$ of (37) agrees with the scaling of (38) for all $\tau \in (2, 3)$. This indicates that a power-law distribution asymptotically achieves the most extreme values of $a(h)$ possible for h small among all random graphs with the same mean and variance of the degree distribution. For larger h, the extremal $a(h)$ scaling is only attained for power-law random graphs for

$\tau = 2$. Indeed, Figure 5a illustrates that the variance-based upper bound on $a(k)$ is close to the value achieved by a power-law Chung-Lu model when $\tau \approx 2$, and that power-law graphs with higher degree-exponents have a closer gap with the maximal possible $a(h)$ value.

We now again compare $a(h)$ of the the power-law distribution (37) to a matching extremal value, but now the extremal random graph based on a matching mean and MAD. For power-laws [44],

$$d = \frac{C(2\mu^{2-\tau} - 1 - h_c^{2-\tau}))}{\tau - 2} + \frac{C\mu(-2\mu^{1-\tau} + 1 + h_c^{1-\tau}))}{\tau - 1}. \tag{40}$$

Thus, for $\tau > 2$, d is approximately constant. Comparing Theorem VI.1 where we take d constant and $h_c = n^{1/(\tau-1)}$ with (37) shows that

$$\max_{\mathbb{P}\in\mathcal{P}^*(\mu,d)} \mathbb{E}_{\mathbb{P}}\left[a(h)\right] = \begin{cases} \frac{d}{2\mu} n^{1/(\tau-1)} & h < \mu n^{(\tau-2)/(\tau-1)} \\ \frac{dn}{2h} & h > \mu n^{(\tau-2)/(\tau-1)}. \end{cases} \tag{41}$$

Comparing this with (37), shows that for $\tau = 2$, the maximal degree-degree correlations among all Chung-Lu random graphs with given MAD, and μ scales the same as for the power-law distribution. Still, Figure 5b shows that the differences in constants in the change point as well as in the maximal scaling make the power-law $a(h)$ to be quite far from the MAD-based bound, even for $\tau \approx 2$. Note also that the MAD-based bounds sometimes drops below the actual power-law value for h large. This is because we plot the values of $a(k)$ all the way up to $h = n$, while the (τ-dependent) cutoff lies already at $h_c = n^{1/(\tau-1)}$, which is close to 400 for $\tau = 2.9$ for example. Thus, for all h in $[1, h_c]$, the MAD-bound is a valid upper bound.

d. Comparing power-law $c(h)$ to extremal $c(h)$. We now turn to $c(k)$. In power-law Chung-Lu models with cutoff $h_c = n^{1/(\tau-1)}$ [41],

$$c(h) \sim \begin{cases} n^{2-\tau}\ln(n) & h \ll n^{(\tau-2)/(\tau-1)} \\ n^{2-\tau}\ln(n/h) & n^{(\tau-2)/(\tau-1)} \ll h \ll \sqrt{n} \\ h^{2\tau-6} n^{5-2\tau} & h \gg \sqrt{n}. \end{cases} \tag{42}$$

We now compare this scaling to the scaling obtained by Theorem IV.1. Using (8) shows that Theorem IV.1 predicts that

$$\max_{\mathbb{P}\in\mathcal{P}(\mu,n^{(3-\tau)/(\tau-1)})} \mathbb{E}_{\mathbb{P}}\left[c(h)\right] \sim \min(n^{(7-3\tau)/(\tau-1)}, 1) \tag{43}$$

for h small, while it scales as n/h^2 for h large. For $\tau = 2$, this coincides with (42). Therefore, Theorem IV.1 implies that the maximal $c(h)$ scaling among all Chung-Lu random graphs with given σ^2 and μ for $h \gg \sqrt{n}$ is achieved by the power-law distribution for $\tau = 2$. However, for $h \ll \sqrt{n}$, the power-law distribution does not attain the largest possible value of $c(h)$. Figure 5c illustrates this. The power-law $c(k)$ is even in its constant, very close to the variance-based maximal value of $c(k)$. For $\tau \approx 2$, the changing point between the constant regime and the decaying regime becomes close, while for larger values of τ, the difference in changing point is larger.

For the MAD-based optimizer, we can obtain similar statements. Theorem VI.2 predicts that

$$\max_{\mathbb{P}\in\mathcal{P}^*(\mu,d)} \mathbb{E}_{\mathbb{P}}\left[c(h)\right] \sim n/h^2 \tag{44}$$

for h sufficiently high. Therefore, Theorem VI.2 implies that the maximal $c(h)$ scaling among all Chung-Lu random graphs with given MAD, h_c and μ for $h \gg \sqrt{n}$ is achieved by the power-law distribution for $\tau = 2$. Indeed, Figure 5d shows that the MAD-based optimal value of $c(k)$ is close to the power-law based one for $\tau \approx 2$, but that the bound can be far off otherwise. Furthermore, note that for $\tau = 2.1$ the MAD-based optimal bound drops below the power-law achieved value. This is because of finite-size effects that are not included in Theorem VI.2.

e. Comparing power-law clique counts to extremal clique counts. We now compare the maximal amount of cliques predicted by Theorem V.1 to the amount of cliques achieved by a power-law random graph. When $2 < \tau < 3$, under a cutoff at $b = \sqrt{\mu n}$, the expected number of cliques in a power-law random graph with degree-exponent τ equals [25, Eq. (1.7)]

$$\mathbb{E}_{pl}[N_{K_k}] \approx \frac{n^{k/2(3-\tau)}\mu^{k/2(1-\tau)}}{k!}\left(\frac{C}{k-\tau}\right)^k. \tag{45}$$

Now under a cutoff at $b = \sqrt{\mu n}$, for power-law random graphs, $\sigma^2 = \frac{C}{3-\tau}\sqrt{\mu n}^{3-\tau}$. Plugging in $\sigma^2 = \frac{C}{3-\tau}\sqrt{\mu n}^{3-\tau}$ from the power-law distribution into (28) yields

$$\max_{\mathbb{P}\in\mathcal{P}^*_{(\mu,\sigma^2)}} \mathbb{E}_{\mathbb{P}}\left[N_{K_k}\right] = \frac{n^{k/2(3-\tau)}\mu^{k/2(1-\tau)}}{k!}\left(\frac{C}{3-\tau}\right)^k. \tag{46}$$

This agrees in terms of scaling in n and μ with the power-law Chung-Lu number of cliques scaling of (45). So this proves that for variance-based clique optimization, power-laws contain the most number of cliques in terms of scaling in n when using a cutoff at $b = \sqrt{\mu n}$. Still, the leading constant in (46) is higher than the one in (45) for $k > 3$, so that the extremal random graph for cliques still achieves a higher total number of cliques in the leading order constant.

When $h_c = n^{1/(\tau-1)}$, then $\mathbb{E}_{pl}[K_k] \sim n^{k(3-\tau)/2}$. Furthermore, in that setting, $\sigma^2 \sim n^{(3-\tau)/(\tau-1)}$, while μ does not grow in n. Therefore, in this case, Theorem V.1 does not apply for $\tau < 7/3$, as $\sigma^2 \geq \sqrt{n}$ when $\tau < 7/3$, so that the condition of Theorem V.1 does not apply. For $\tau > 7/3$, plugging in the power-law value of $\sigma^2 \sim n^{(3-\tau)/(\tau-1)}$ into Theorem V.1 yields

$$\max_{\mathbb{P}\in\mathcal{P}^*(\mu,\sigma^2)} \mathbb{E}_{\mathbb{P}}\left[N(K_k)\right] \sim n^{k(3-\tau)/(\tau-1)}, \tag{47}$$

which is larger than the power-law scaling of $n^{k(3-\tau)/2}$. Therefore, power-law random graphs are not the graphs that contain the most cliques among all random graphs with the same mean and variance.

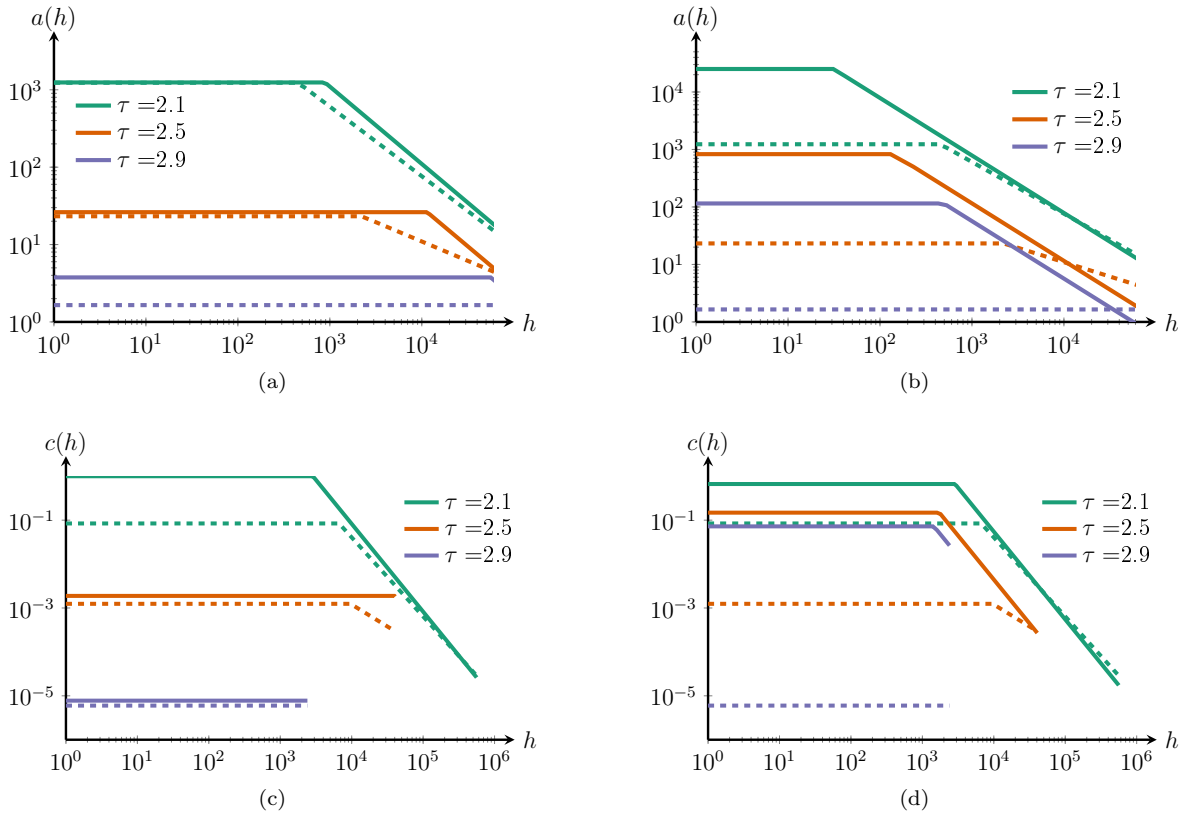


FIG. 5. Maximal scaling compared with power-law with the same parameters (dashed line) for $n = 10^5$ on a) $a(k)$, variance basis (solid line), b) $a(k)$, MAD-basis (solid line) c) $c(k)$, variance basis (solid line), d) $c(k)$, MAD-basis (solid line).

f. Data We now apply our bounds for $a(k)$ and $c(k)$ to three real-world network data sets. Figure 6 compares the variance-based upper bound of Theorem III.1 with empirical observations. For all data sets, the true value of $a(k)$ exceeds the variance-based maximizer at some point. The MAD-based maximizer on the other hand, remains an upper bound for $a(k)$ in almost all data sets. This highlights the importance of the right choice of comparison model: The hidden-variable model with fitted variance cannot explain the degree-degree correlations in these data sets, while the same model with fitted MAD can.

Figure 7 shows for three real-world networks $c(k)$ and the variance- and MAD-based bounds. The $c(k)$-values of the Gowalla data set are close to, or below the MAD and the variance-based optimizer, respectively. This suggests that these data sets can be suitably modeled by some hidden-variable model that matches the $c(k)$ distribution of this data set. For the two other data sets on the other hand, the value of $c(k)$ in the data sets is higher than can be achieved by any hidden-variable model. Therefore, no hidden-variable model is able to match these data sets in terms of $c(k)$, which is likely caused by the locally-tree like nature of the hidden-variable model.

VIII. CONCLUSIONS AND DISCUSSION

Our robust network perspective, in terms of ambiguity and partial information on the degree distribution, comes with substantial mathematical challenges. We created an optimization framework for identifying, within some ambiguity set, the extreme degree distribution that generates the upper bound for the degree-degree correlations and clustering. We therefore had to combine probabilistic models (random graphs) with optimization models (stochastic programs). For successfully applying our robust perspective it is crucial to solve a given stochastic program in closed form. Here we distinguish between 1D programs such as the semi-infinite LP for the expected degree-degree correlation, and 2D programs such as for the expected clustering. For the 1D programs with both variance and MAD information, we were able to apply standard primal-dual techniques, solving the dual problem with the tightest majorant, and indeed finding a closed-form extremal distribution.

The 2D programs for expected clustering proved to be more challenging, being semi-infinite programs with two i.i.d. random variables. This i.i.d. assumption creates nonlinear conditions that prohibits the usage of standard primal-dual techniques. For variance information, we therefore applied a relaxation technique that replaces

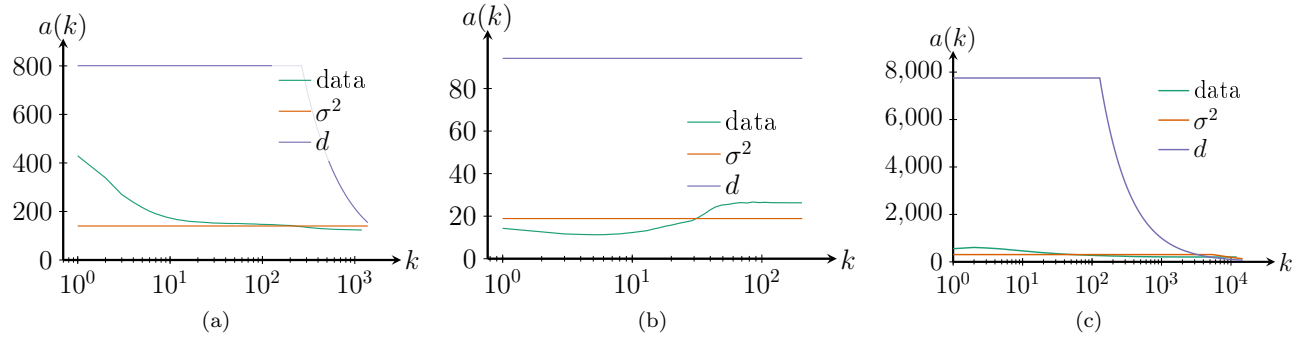


FIG. 6. $a(k)$ and MAD and variance-based bounds for 3 real-world networks a) Enron email network [27], b) Pretty Good Privacy network [5], c) Gowalla social network [11].

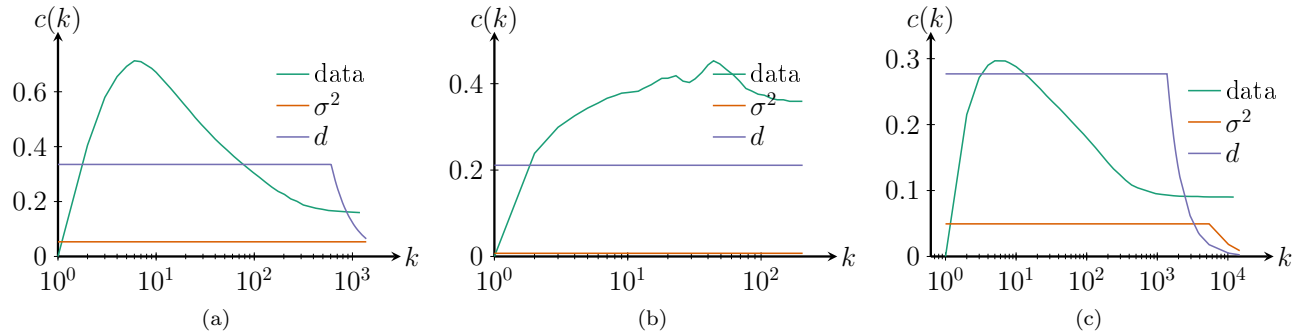


FIG. 7. $c(k)$ and MAD and variance-based bounds for 3 real-world networks, a) Enron email network [27], b) Pretty Good Privacy network [5], c) Gowalla social network [11].

the original program with its counterpart that allows correlations. That relaxed program proved solvable with the primal-dual technique, despite the additional challenges of 2D instead of 1D. Surprisingly, the found extremal joint distribution was a product-form distribution after all, so that we could thus show that the relaxed program has the same solution as the original program. We showed that similar relaxations for programs of higher dimensions could be also be used to establish tight (asymptotic) bounds for clique counts.

However, such relaxations proved cumbersome, if not intractable, for MAD information. In that case, we opted for a different relaxation, in which we solve the 2D program in two steps: first finding the worst-case distributions of variable 1, and then, given this worst-case distribution, finding the worst-case distribution distribution of variable 2. This relaxation also allows correlation between the random variables, but possibly of a different nature. This second type of relaxation turned out to give worst-case distributions that become independent (product-form) distributions when the network size grows to infinity. Hence, in this way, we could solve the original 2D MAD program asymptotically for $n \to \infty$.

This paper forms an important step towards a more complete theory of Distributionally Robust Random Graphs (DRRGs). This theory exchanges full information random graphs with partial information models, for instance regarding the degree distribution. This means that the theory applies to a large class of distributions, and can possibly explain complex network phenomena in a more universal manner, less dependent on the specific distributional assumptions. Hence, we consider this a robust way of studying complex networks.

Here we mention a few open problems and research directions. For solving the 2D program we have introduced two relaxations, one for variance and one for MAD. Do both relaxations work for both variance and MAD? Can we understand when the two relaxations are equivalent, and when they are not? Do these relaxations also work in higher-dimensional stochastic programs, with more than two i.i.d. random variables? In this paper we have successfully applied one such higher-dimensional relaxation for cliques of arbitrary size.

Another avenue for future research relates to model extension. Our data analysis revealed several data sets whose $c(k)$ cannot be matched by any hidden-variable model. This motivates to seek for comparable robust bounds for other random graph models, such as the GIRG [7], which is a generalization of the popular hyperbolic random graph [28] where the connection probability of two vertices scales as the product of their weights, divided by their distance, or a random geometric

graph [35]. What is the maximal value of $c(k)$ for given mean and variance on the degrees and given mean and variance on the inter-distances? This question will lead to a more involved optimization problem with more variables due to the underlying geometry. Such upper bounds for increasing model complexity could detect what level of complexity is necessary to model a specific network property correctly.

Finally, while we investigated robust degree distributions, one can also think of other network properties. For example, in temporal network models one can obtain results for robust edge time-stamps. For hypergraphs, one can think of robust hyperdegrees, or of robust positions for geometric models. We believe that this framework can provide robust upper bounds on several network properties, and quantify the sensitivity of network models to specific assumptions on their parameters.

ACKNOWLEDGMENTS

We thank Dick den Hertog, Wouter van Eekelen and Pieter Kleer for various thought exchanges about the relaxations introduced in this paper for solving semi-infinite linear programs, and robust optimization in general. JB is supported by an NWO Mathematics Clusters grant, JvL is supported by an NWO VICI grant. CS is supported by NWO VENI grant 202.001 and NWO M2 grant 0.379.

[1] B.-T. A. and E. Hochman. More bounds on the expectation of a convex function of a random variable. *Journal of Applied Probability*, 9(4):803–812, dec 1972.

[2] G. Bianconi and M. Marsili. Loops of any size and hamilton cycles in random scale-free networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(06):P06005, 2005.

[3] T. Blasius, T. Friedrich, A. Krohmer, and S. Laue. Efficient embedding of scale-free graphs in the hyperbolic plane. *IEEE/ACM Transactions on Networking*, 26(2):920–933, 2018.

[4] M. Boguñá and R. Pastor-Satorras. Class of correlated random networks with hidden variables. *Phys. Rev. E*, 68:036112, 2003.

[5] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas. Models of social networks based on social distance attachment. *Phys. Rev. E*, 70(5), 2004.

[6] M. Boguná, R. Pastor-Satorras, and A. Vespignani. Cutoffs and finite size effects in scale-free networks. *The European Physical Journal B*, 38(2):205–209, 2004.

[7] K. Bringmann, R. Keusch, and J. Lengler. Geometric inhomogeneous random graphs. *Theoretical Computer Science*, 760:35–54, feb 2019.

[8] A. D. Broido and A. Clauset. Scale-free networks are rare. *Nature Communications*, 10(1), mar 2019.

[9] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Phys. Rev. E*, 71:027103, Feb 2005.

[10] L. Chen, D. Padmanabhan, C. C. Lim, and K. Natarajan. Correlation robust influence maximization. *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*, Oct. 2020.

[11] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[12] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci. USA*, 99(25):15879–15882, 2002.

[13] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, 2009.

[14] P. Colomer-de Simon and M. Boguñá. Clustering of random scale-free networks. *Phys. Rev. E*, 86:026120, 2012.

[15] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 66(3), sep 2002.

[16] W. van Eekelen, D. den Hertog, and J. S. H. van Leeuwaarden. MAD dispersion measure makes extremal queue analysis simple. 2019.

[17] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pages 251–262. ACM, 1999.

[18] R. J. Gallagher, J.-G. Young, and B. F. Welles. A clarified typology of core-periphery structure in networks. *Science Advances*, 7(12), mar 2021.

[19] S. Han, M. Tao, U. Topcu, H. Owhadi, and R. M. Murray. Convex optimal uncertainty quantification. *SIAM Journal on Optimization*, 25(3):1368–1387, 2015.

[20] R. van der Hofstad, A. J. E. M. Janssen, J. S. H. van Leeuwaarden, and C. Stegehuis. Local clustering in scale-free networks with hidden variables. *Phys. Rev. E*, 95(2):022307, 2017.

[21] R. van der Hofstad, J. S. H. van Leeuwaarden, and C. Stegehuis. Optimal subgraph structures in scale-free configuration models. *The Annals of Applied Probability*, 31(2):501–537, 2021.

[22] K. Isii. On sharpness of Tchebycheff-type inequalities. *Annals of the Institute of Statistical Mathematics*, 14(1):185–197, 1962.

[23] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon. Subgraphs in random networks. *Physical review E*, 68(2):026127, 2003.

[24] S. Janson. On percolation in random graphs with given vertex degrees. *Electron. J. Probab.*, 14:86–118, 2009.

[25] A. J. E. M. Janssen, J. S. H. van Leeuwaarden, and S. Shneer. Counting cliques and cycles in scale-free inhomogeneous random graphs. *Journal of Statistical Physics*, 175(1):161–184, feb 2019.

[26] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[27] B. Klimt and Y. Yang. Introducing the Enron Corpus. In *CEAS*, 2004.

[28] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguná. Hyperbolic geometry of complex networks. *Phys. Rev. E*, 82(3):036106, 2010.

[29] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, 2014. Date of access: 14/03/2017.

[30] S. Mossa, M. Barthélémy, H. E. Stanley, and L. A. N. Amaral. Truncation of power law behavior in "scale-free" network models due to information filtering. *Physical Review Letters*, 88(13), mar 2002.

[31] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, jan 2001.

[32] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64(2):026118, 2001.

[33] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the internet. *Phys. Rev. Lett.*, 87:258701, 2001.

[34] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200, 2001.

[35] M. Penrose. *Random Geometric Graphs*. Oxford University Press, may 2003.

[36] I. Popescu. A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operations Research*, 30(3):632–657, 2005.

[37] T. A. Ramasubban. The mean difference and the mean deviation of some discontinuous distributions. *Biometrika*, 45(3/4):549, dec 1958.

[38] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67:026112, 2003.

[39] W. W. Rogosinski. Moments of non-negative mass. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 245(1240):1–27, 1958.

[40] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[41] C. Stegehuis. Degree correlations in scale-free random graph models. *Journal of Applied Probability*, 56(3):672–700, 2019.

[42] C. Stegehuis, R. van der Hofstad, and J. S. H. van Leeuwaarden. Variational principle for scale-free network motifs. *Scientific Reports*, 9(1):6762, 2019.

[43] C. Stegehuis, R. van der Hofstad, J. S. H. van Leeuwaarden, and A. J. E. M. Janssen. Clustering spectrum of scale-free networks. *Phys. Rev. E*, 96(4):042309, 2017.

[44] J. S. H. van Leeuwaarden and C. Stegehuis. Robust subgraph counting with distribution-free random graph analysis. *Physical Review E*, 104(4):044313, oct 2021.

[45] A. Vázquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of the internet. *Phys. Rev. E*, 65:066130, 2002.

[46] I. Voitalov, P. van der Hoorn, R. van der Hofstad, and D. Krioukov. Scale-free networks well done. *Physical Review Research*, 1(3):033034, 2019.

[47] D. Yao, P. van der Hoorn, and N. Litvak. Average nearest neighbor degrees in scale-free networks. *Internet Mathematics*, 2018.

Appendix A: Proofs

1. Proof of the variance-based maximizer for $c(k)$ and the number of cliques

Proof of Theorem IV.1. The optimization problem (23) is equivalent to

$$\begin{aligned} \max_{\mathbb{P}(x)\geq 0} \quad & \int_{x_1}\int_{x_2} g(x_1,x_2)\mathrm{d}\mathbb{P}(x_2)\mathrm{d}\mathbb{P}(x_1) \\ \text{s.t.} \quad & \int_{x_1}\int_{x_2} x_1^2x_2^2\mathrm{d}\mathbb{P}(x_2)\mathrm{d}\mathbb{P}(x_1) = (\mu^2+\sigma^2)^2 \\ & \int_{x_1}\int_{x_2} x_1x_2\mathrm{d}\mathbb{P}(x_2)\mathrm{d}\mathbb{P}(x_1) = \mu^2, \\ & \int_{x_1}\int_{x_2} \mathrm{d}\mathbb{P}(x_2)\mathrm{d}\mathbb{P}(x_1) = 1. \end{aligned} \tag{A1}$$

We now consider a relaxed version of this optimization problem. Instead of drawing from a single measure $\mathbb{P}$ for both x_1 and x_2, we allow for a dual measure $\mathbb{P}(x_1,x_2)$, where we only require the product of the means and second moments to be equal to μ^2 and $(\mu^2+\sigma^2)^2$, respectively. We thus drop the i.i.d. assumption for now. This gives the problem

$$\begin{aligned} \max_{\mathbb{P}(x_1,x_2)\geq 0} \quad & \int_{x_1}\int_{x_2} g(x_1,x_2)\mathrm{d}\mathbb{P}(x_1,x_2) \\ \text{s.t.} \quad & \int_{x_1}\int_{x_2} x_1^2x_2^2\mathrm{d}\mathbb{P}(x_1,x_2) = (\mu^2+\sigma^2)^2 \\ & \int_{x_1}\int_{x_2} x_1x_2\mathrm{d}\mathbb{P}(x_1,x_2) = \mu^2, \\ & \int_{x_1}\int_{x_2} \mathrm{d}\mathbb{P}(x_1,x_2) = 1. \end{aligned} \tag{A2}$$

The dual problem then becomes

$$\begin{aligned} \min_{\lambda_1,\lambda_2,\lambda_3} \quad & \lambda_1(\mu^2+\sigma^2)^2 + \lambda_2\mu^2 + \lambda_3 \\ \text{s.t.} \quad & g(x_1,x_2) - \lambda_1x_1^2x_2^2 - \lambda_2x_1x_2 - \lambda_3 \leq 0 \\ & \forall x_1,x_2 \in [a,h_c], \end{aligned}$$

with

$$g(x_1,x_2) = \min\Big(\frac{x_1x_2}{\mu n},1\Big)\min\Big(\frac{x_1h}{\mu n},1\Big)\min\Big(\frac{x_2h}{\mu n},1\Big). \tag{A3}$$

We will now solve the optimization problem by constructing a primal and dual solution that achieve the same objective value, and therefore optimize (A2). Furthermore, the constructed optimal probability distribution turns out to be of product form, so that they must also be optimizers of the original, more constrained optimization problem (A1). These primal and dual solutions depend on h, n, μ and σ, in the following cases:

Case 1: $h \leq \sqrt{\mu n}$. We take the dual solution $\lambda_1 = h^2/(\mu n)^3$, $\lambda_2 = \lambda_3 = 0$. This gives as objective value $(\mu^2+\sigma^2)^2h^2/(\mu n)^3$.

When $\sigma^2 \leq (\sqrt{\mu n} - \mu)\mu$, for the primal problem, consider the 3-point distribution

$$p_0 = \frac{\sigma^2}{\sqrt{\mu n}\mu}, \quad p_\mu = 1 - \frac{\sigma^2}{\mu(\sqrt{\mu n} - \mu)},$$
$$p_{\sqrt{\mu n}} = \frac{\sigma^2}{\sqrt{\mu n}(\sqrt{\mu n} - \mu)}. \tag{A4}$$

This is a proper distribution by the condition on σ^2. Then,

$$\begin{aligned}
&\mathbb{E}\left[g(X_1, X_2)\right] \\
&= \frac{h^2}{(\mu n)^3}\left(p_\mu^2 \mu^4 + 2p_\mu p_{\sqrt{\mu n}} \mu^2(\mu n) + p_{\sqrt{\mu n}}^2 (\mu n)^2\right) \\
&= \frac{h^2}{(\mu n)^3}\left(p_\mu \mu^2 + p_{\sqrt{\mu n}} \mu n\right)^2 = \frac{h^2}{(\mu n)^3}(\mu^2 + \sigma^2)^2.
\end{aligned} \tag{A5}$$

Thus, by strong duality, this is the optimizer of (23).

When $\sigma^2 > (\sqrt{\mu n} - \mu)\mu$, consider the three-point distribution

$$p_0 = 1 - p_{\sqrt{\mu n}} - p_{\mu n/h},$$
$$p_{\sqrt{\mu n}} = \frac{\mu^2 + \sigma^2 - \mu n/h \cdot \mu}{\sqrt{\mu n}(\sqrt{\mu n} - \mu n/h)},$$
$$p_{\mu n/h} = \frac{\mu^2 + \sigma^2 - \sqrt{\mu n} \cdot \mu}{\mu n/h(\mu n/h - \sqrt{\mu n})}. \tag{A6}$$

This is only a proper distribution when $\sigma^2 < (\mu n/h - \mu)\mu$. This three-point distribution gives $\mathbb{E}\left[c(h)\right] = 1$, so that it achieves the maximum $c(h)$. Therefore, we can immediately conclude that this primal solution is optimal.

Case 2: $\sqrt{\mu n} < h < \mu^2 n/(\mu^2 + \sigma^2)$. We take as dual solution again $\lambda_1 = h^2/(\mu n)^3$, $\lambda_2 = \lambda_3 = 0$. This gives as objective value $(\mu^2 + \sigma^2)^2 h^2/(\mu n)^3$.

For the primal problem, consider the 3-point distribution

$$p_0 = \frac{\sigma^2}{\mu(\mu n/h)}, \quad p_\mu = 1 - \frac{\sigma^2}{\mu(\mu n/h - \mu)},$$
$$p_{\mu n/h} = \frac{\sigma^2}{\mu n/h(\mu n/h - \mu)}. \tag{A7}$$

Again, this is only a distribution when $\sigma^2 \leq \mu(\mu n/h - \mu)$, which is ensured by the condition on h. Then,

$$\begin{aligned}
&\mathbb{E}\left[g(X_1, X_2)\right] \\
&= \frac{h^2}{(\mu n)^3}\left(p_\mu^2 \mu^4 + 2p_\mu p_{\mu n/h} \mu^2(\mu n)^2/h^2 + p_{\mu n/h}^2 (\mu n/h)^4\right) \\
&= \frac{h^2}{(\mu n)^3}\left(p_\mu \mu^2 + p_{\mu n/h}(\mu n/h)^2\right)^2 = \frac{h^2}{(\mu n)^3}(\mu^2 + \sigma^2)^2.
\end{aligned} \tag{A8}$$

Thus, the primal solution achieves the same value as the dual solution. Therefore, by strong duality, this is the optimizer of (23).

Case 3: $h \geq \mu^2 n/(\mu^2 + \sigma^2)$. We take as dual solution $\lambda_2 = 1/(\mu n)$, $\lambda_1 = \lambda_3 = 0$. This gives as objective value μ/n.

For the primal problem, consider again the 3-point distribution that is given in (A6). This is only a proper distribution when $\sigma^2 > (\mu n/h - \mu)\mu$, which is satisfied by our condition on σ^2, and $\sigma^2 \leq (\sqrt{\mu n} - \mu)\mu$. Under this three-point distribution, $\mathbb{E}[X] = \mu$, $\mathbb{E}[(X-\mu)^2] = \sigma^2$ and $\mathbb{E}[g(X_1, X_2)] = \mu/n$. Thus, by strong duality, this is the optimal solution. □

Proof of Theorem V.1. The relaxed optimization problem corresponding to (27) gives the dual problem

$$\begin{aligned}
\min_{\lambda_1, \lambda_2, \lambda_3} \quad & \lambda_1(\mu^2 + \sigma^2)^k + \lambda_2 \mu^k + \lambda_3 \\
\text{s.t.} \quad & g(x_1, \ldots, x_k) - \lambda_1 x_1^2 \cdots x_k^2 - \lambda_2 x_1 x_2 \cdots x_k - \lambda_3 \leq 0 \\
& \forall x_1, \ldots, x_k \in [a, h_c],
\end{aligned}$$

with

$$g(x_1, \ldots, x_k) = \prod_{1 \leq i < j \leq k} \min\left(\frac{x_i x_j}{\mu n}, 1\right). \tag{A9}$$

Consider the dual solution $\lambda_1 = 1/(\mu n)^k$, giving as objective value $(\mu^2 + \sigma^2)^k/(\mu n)^k$.

Consider the 3-point distribution

$$p_0 = 1 - p_m - p_{\sqrt{\mu n}}, \quad p_m = \frac{\mu\sqrt{\mu n} - \mu^2 - \sigma^2}{m(\sqrt{\mu n} - m)},$$
$$p_{\sqrt{\mu n}} = \frac{\mu^2 + \sigma^2 - m\mu}{\sqrt{\mu n}(\sqrt{\mu n} - m)}, \tag{A10}$$

with $m = \mu^{(1+k)/4} n^{(3-k)/4}$. Note that $m < \sqrt{\mu n}$ for $k > 3$, and $m = \mu$ for $k = 3$, and that this is only a proper distribution when $\sigma^2 \leq \mu(\sqrt{\mu n} - \mu)$. Now

$$\mathbb{E}\left[g(X_1, \ldots, X_k)\right] = \frac{\mathbb{E}\left[X_1^{k-1}\right]^k}{(\mu n)^{k(k-1)/2}}. \tag{A11}$$

Furthermore,

$$\begin{aligned}
\mathbb{E}\left[X_1^{k-1}\right] = &\frac{\left(\mu^{\frac{k+1}{4}} n^{\frac{3-k}{4}}\right)^{k-2}\left(\mu\sqrt{\mu n} - \mu^2 - \sigma^2\right)}{\sqrt{\mu n} - \mu^{\frac{k+1}{4}} n^{\frac{3-k}{4}}} \\
&+ \frac{\sqrt{\mu n}^{k-2}\left(\mu^2 + \sigma^2 - \mu^{\frac{k+5}{4}} n^{\frac{3}{4} - \frac{k}{4}}\right)}{\sqrt{\mu n} - \mu^{\frac{k+1}{4}} n^{\frac{3-k}{4}}}.
\end{aligned} \tag{A12}$$

Now when $k > 3$, then $n^{(3-k)/4} = o(1)$. Therefore, for $k > 3$,

$$\mathbb{E}\left[X_1^{k-1}\right] = \frac{\sqrt{\mu n}^{k-2}\left(\mu^2 + \sigma^2\right)}{\sqrt{\mu n}}(1 + o(1)). \tag{A13}$$

Thus, also

$$\begin{aligned}
\mathbb{E}\left[g(X_1, \ldots, X_k)\right] &= \frac{\sqrt{\mu n}^{(k-3)k}(\mu^2 + \sigma^2)^k}{(\mu n)^{k(k-1)/2}}(1 + o(1)) \\
&= \frac{(\mu^2 + \sigma^2)^k}{(\mu n)^k}(1 + o(1)),
\end{aligned} \tag{A14}$$

making the 3-point distribution asymptotically optimal, as it asymptotically achieves the same value as the dual solution. Furthermore, for $k = 3$, $\mathbb{E}\left[X^2\right] = \mu^2 + \sigma^2$, by the conditions in $\mathcal{P}(\mu, \sigma)$. Thus, for $k = 3$,

$$\mathbb{E}\left[g(X_1, \ldots, X_k)\right] = \frac{(\mu^2 + \sigma^2)^3}{(\mu n)^3}, \tag{A15}$$

which is the exact same value as the dual objective value. Hence, by strong duality, this is the optimal solution. □

2. Proofs for MAD-based maximizers of $a(k)$ and $c(k)$

Proof of Theorem VI.1. The function $h' \min(\frac{hh'}{\mu n}, 1)$ is piecewise convex in h'. In particular, it is quadratic up to $l = \min(\mu n/h, h_c) \gg \mu$, where it has slope 2, and it is linear with slope 1 after that.

Thus, to optimize over the distribution of h', we need to solve

$$\begin{aligned} &\max_{\mathbb{P} \in \mathcal{P}(\mu, d)} \int_x f(x) \mathrm{d}\mathbb{P}(x) \\ &\text{s.t.} \quad \int_x |x - \mu| \mathrm{d}\mathbb{P}(x) = d, \int_x x \mathrm{d}\mathbb{P}(x) = \mu, \int_x \mathrm{d}\mathbb{P}(x) = 1, \end{aligned} \tag{A16}$$

where $f(x) = x \min(\frac{hx}{\mu n}, 1)$. Similarly to the derivation of [16, Eq. (6)], this results in the dual problem

$$\begin{aligned} &\min_{\lambda_1, \lambda_2, \lambda_3} \lambda_1 d + \lambda_2 \mu + \lambda_3 \\ &\quad \text{s.t. } f(x) - \lambda_1 |x - \mu| - \lambda_2 x - \lambda_3 \leq 0 \quad \forall x \in [0, h_s]. \end{aligned} \tag{A17}$$

For simplicity of notation, we assume that $a = 0$. Thus, this dual problem aims to find the tightest piecewise linear majorant of $f(x)$ with a kink at μ that minimizes the objective value. Consider the majorant $F_1(x) = \frac{hl}{2\mu n}|x - \mu| + (\frac{hl}{2\mu n} + \frac{h}{n})x - \frac{hl}{2n}$. Now $F_1(x)$ has as its objective value

$$\frac{hl}{2\mu n} d + (\frac{hl}{2\mu n} + \frac{h}{n})\mu - \frac{hl}{2n} = \frac{\mu h}{n} + \frac{hdl}{2\mu n}. \tag{A18}$$

By weak duality of semi-infinite linear programming, we know that a feasible solution to the dual problem provides us with a valid upper bound for the optimal primal solution value. Thus, we now find a feasible primal solution with an objective value equal to this upper bound results to achieve strong duality. As the tightest majorant of the dual problem touches $f(x)$ at the points a, μ and $l = \min(\mu n/h, h_c)$, we consider the three-point distribution

$$\begin{aligned} p_0 &= \frac{d}{2\mu}, \quad p_\mu = 1 - \frac{d}{2\mu} - \frac{d}{2(l - \mu)}, \\ p_l &= \frac{d}{2(l - \mu)}, \end{aligned} \tag{A19}$$

which is a distribution since $\mu n/h \gg \mu$. This yields as objective value for the primal problem

$$\frac{\mu^2 h}{\mu n}\Big(1 - \frac{d}{2\mu} - \frac{d}{2(l - \mu)}\Big) + \frac{l^2 h}{\mu n}\frac{d}{2(l - \mu)} = \frac{\mu h}{n} + \frac{hdl}{2\mu n}. \tag{A20}$$

Thus, we have strong duality as the primal objective from (A20) and the dual optimal value are the same.

Therefore,

$$\mathbb{E}_{\mathbb{P}}\left[a(h)\right] = \frac{n}{h}\Big(\frac{\mu h}{n} + \frac{hdl}{2\mu n}\Big) = \mu + \frac{dl}{2\mu}. \tag{A21}$$

Now when h_c tends to infinity the second term dominates as $l = \min(h_c, \mu n/h) \gg 1$ when $h \ll n$ and,

$$\mathbb{E}_{\mathbb{P}}\left[a(h)\right] = \frac{d}{2\mu} \min(h_c, \frac{\mu n}{h})(1 + o(1)). \tag{A22}$$

□

Proof of Theorem VI.2. Case 1: $h \ll \sqrt{\mu n}$ and $2\mu(1 - \sqrt{\mu/n}) < d < 2\mu(1 - h/n)$. In this case, consider the three-point distribution

$$\begin{aligned} p_0 &= \frac{d}{2\mu}, \quad p_{\sqrt{\mu n}} = \frac{(d - 2\mu)l + 2\mu^2}{2\mu\left(\sqrt{\mu n} - l\right)}, \\ p_l &= \frac{(d - 2\mu)\sqrt{\mu n} + 2\mu^2}{2\mu\left(l - \sqrt{\mu n}\right)}, \end{aligned} \tag{A23}$$

for $l = \mu n/h$, which is a proper distribution under the condition $2\mu(1 - \sqrt{\mu/n}) < d < 2\mu(1 - h/n)$. For this three-point distribution,

$$\mathbb{E}_{\mathbb{P}}\left[c(h)\right] = \frac{n^2}{h^2}\Big(p^2_{\sqrt{\mu n}}\frac{\mu n h^2}{(\mu n)^2} + 2p_l p_{\sqrt{\mu n}}\frac{\sqrt{\mu n} h}{\mu n} + p_l^2\Big) = 1. \tag{A24}$$

As $c(h)$ is a probability, we have $\mathbb{E}\left[c(h)\right] \leq 1$, so that it coincides with the upper bound.

Case 2: $h \gg \sqrt{\mu n}$. We now apply the three-point optimization problem in two steps: first for the optimal distribution of h_i, then for h_j. We will show that these two optimal distributions are identical and independent, so that this method shows that optimizing the distribution of h_i and h_j while constraining them to be equal yields this same optimal distribution.

The function we would like to optimize is

$$\mathbb{E}\left[\min(\frac{hh_1}{\mu n}, 1) \min(\frac{hh_2}{\mu n}, 1) \min(\frac{h_1 h_2}{\mu n}, 1)\right]. \tag{A25}$$

Step 1: optimizing over h_1. If we optimize only over the distribution of h_1 and fix h_2, this is equivalent to optimizing

$$\mathbb{E}\left[\min(\frac{hh_1}{\mu n}, 1) \min(\frac{h_1 h_2}{\mu n}, 1)\right]. \tag{A26}$$

Thus, we again want to maximize (A16) and therefore minimize its dual problem (A17), but now with

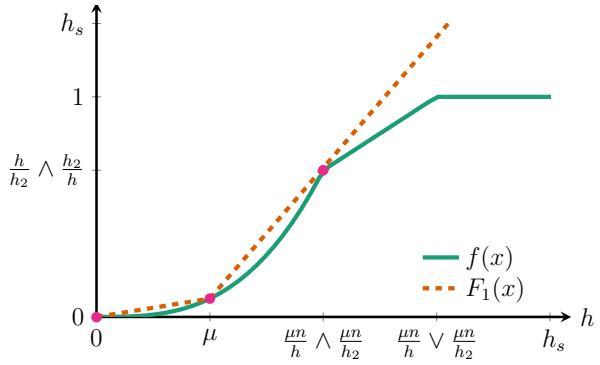


FIG. 8. The tightest majorant for $f(x)$ describing the optimizer over h_1 of $c(h)$

$f(x) = \min(\frac{hx}{\mu n}, 1)\min(\frac{xh_2}{\mu n}, 1)$. We again focus on the dual problem, which is again equivalent to minimizing $\lambda_1 d + \lambda_2 \mu + \lambda_3$ over all tightest majorants of $f(x)$. Now $f(x)$ is again piecewise convex: it is quadratic in h_1 for $h_1 < \min(\mu n/h, \mu n/h_2)$, linear for $h_1 \in [\min(\mu n/h, \mu n/h_2), \max(\mu n/h, \mu n/h_2)]$, and constant for $h_1 \in [\max(\mu n/h, \mu n/h_2), h_c]$. Thus, $f(x)$ is shaped as the function in Figure 8 (where for simplicity $a = 0$).

In the next computations, we assume for simplicity of notation that $a = 0$. Now the tightest majorant of Figure 8, $F_1(x)$, can be parametrized by $\lambda_1 = \min(h, h_2)/(2\mu n)$, $\lambda_2 = \min(h, h_2)/(2\mu n) + hh_2/(\mu n^2)$, $\lambda_3 = -\min(h, h_2)/(2n)$. This gives as objective value for the dual program (A17)

$$d\frac{\min(h,h_2)}{2\mu n} + \mu\Big(\frac{\min(h,h_2)}{2\mu n} + \frac{hh_2}{\mu n^2}\Big) - \frac{\min(h,h_2)}{2n} = \frac{d}{2\mu n}\min(h,h_2) + \frac{hh_2}{n^2}. \tag{A27}$$

We now again consider the primal problem (A16). The solution to the dual problem (A17), $F_1(x)$ has three touching points of $f(x)$: at 0, μ and $\min(\mu n/h_2, \mu n/h)$. Thus, we will now show that $c(h)$ is maximized over h_1 by the three-point distribution

$$p_0 = \frac{d}{2\mu}, \quad p_\mu = 1 - \frac{d}{2\mu} - \frac{d}{2(\min(\mu n/h_2, \mu n/h) - \mu)}, \quad p_{\min(\mu n/h_2, \mu n/h)} = \frac{d}{2(\min(\mu n/h_2, \mu n/h) - \mu)}. \tag{A28}$$

This gives an objective value of (A16) of

$$0 \cdot \frac{d}{2\mu} + \frac{\mu^2 hh_2}{(\mu n)^2}\Big(1 - \frac{d}{2\mu} - \frac{d}{2(\min(\mu n/h_2, \mu n/h) - \mu)}\Big) + \frac{\min(\mu n/h_2, \mu n/h)^2 hh_2}{(\mu n)^2} \cdot \frac{d}{2(\min(\mu n/h_2, \mu n/h) - \mu)} = \frac{d}{2\mu n}\min(h,h_2) + \frac{hh_2}{n^2}. \tag{A29}$$

Thus, the objective value of the three-point distribution is equal to the objective value of the dual problem in (A27). Thus, by strong duality, (A28) is the optimal distribution maximizing (A26).

Step 2: optimizing over h_2. We now plug the optimal three-point distribution of h_1 (A28) into (A25) and then optimize only over the distribution of h_2. We then need to optimize

$$\mathbb{E}\left[(1 - \frac{d}{2\mu} - \frac{d}{2(\min(\mu n/h_2, \mu n/h) - \mu)})\frac{hh_2}{n^2}\min(\frac{hh_2}{\mu n}, 1)\right] + \mathbb{E}\left[\frac{d}{2(\min(\mu n/h_2, \mu n/h) - \mu)}\min(\frac{h}{h_2}, \frac{h_2}{h})\min(\frac{hh_2}{\mu n}, 1)\right]. \tag{A30}$$

This is a function in h_2 that looks like Figure 9: first a convex part, then a linear part, and then again a convex part. The tightest majorant $F_1(x)$ is depicted in Figure 9 as well. $F_1(x)$ is characterized by $\lambda_1 = \tilde{C}h/(2(n\mu)^2)$, $\lambda_2 = \tilde{C}(h^2/(\mu^2 n^3) - h/(2(n\mu)^2))$ and $\lambda_3 = -\tilde{C}h/(2n^2\mu)$, with $\tilde{C} = \mu^2 + dn\mu/(2h)$. This gives as dual objective

$$\lambda_1 d + \lambda_2 \mu + \lambda_3 = \frac{(dn + 2h\mu)^2}{4\mu n^3}. \tag{A31}$$

We now consider the 3-point distribution for h_2 on the touching points $0, \mu$ and $\mu n/h$:

$$p_0 = \frac{d}{2\mu}, \quad p_\mu = 1 - \frac{d}{2\mu} - \frac{d}{2(\mu n/h - \mu)}, \quad p_{\mu n/h} = \frac{d}{2(\mu n/h - \mu)}. \tag{A32}$$

This gives as objective value for the primal problem

$$\mathbb{E}\left[\Big(1 - \frac{d}{2\mu} - \frac{d}{2(\mu n/h - \mu)}\Big)\frac{h^2}{\mu n^3}h_2^2\right] + \mathbb{E}\left[\frac{d}{2(\mu n/h - \mu)}\frac{h_2^2}{\mu n}\right] = \frac{(dn + 2h\mu)^2}{4\mu n^3}, \tag{A33}$$

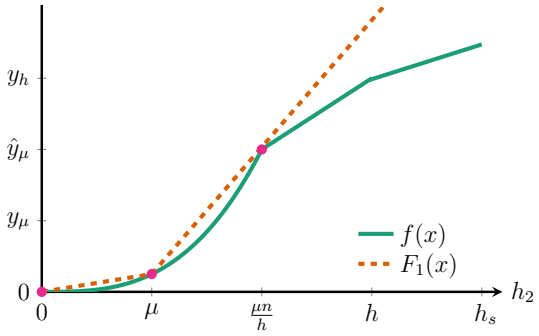


FIG. 9. The tightest majorant for $f(x)$ describing the optimizer over h_2 of $c(h)$

so that by strong duality, this is the optimal three-point distribution for h_2. As $h \gg \sqrt{\mu n}$, this means that $h_2 < h$ for all three values of the three-point distribution. Then, the three-point distribution for (A28) reduces to (A32).

Thus, by optimizing the distributions of h_1 and h_2 separately, we obtain the same three-point distribution for both. Therefore, the optimization of the distributions of h_1 and h_2 where they are constrained to have the same distribution also gives the three-point distribution (A32) as its solution.

Indeed,

$$\max_{x,y:x=y} f(x,y) \leq \max_y \max_x f(x,y), \tag{A34}$$

as all combinations $f(x,x)$ are also encountered on the right-hand side. Furthermore, let x^* and y^* denote the optimizers obtained in the right-hand side, and suppose that $x^* = y^*$. Then,

$$\max_{x,y:x=y} f(x,y) \geq f(x^*,y^*) = \max_y \max_x f(x,y). \tag{A35}$$

Thus, when optimizing the distributions of h_1 and h_2 separately yields an optimizer where both distributions are equal, then this is also the optimization of the distributions of h_1 and h_2 where they are constrained to have the same distribution.

This gives for $c(h)$

$$\begin{aligned}\max_{\mathbb{P}\in\mathcal{P}(\mu,d)} \mathbb{E}_{\mathbb{P}}\left[c(h)\right] &= \frac{n^2}{h^2}\frac{(dn+2h\mu)^2}{4\mu n^3} \\ &= \frac{d^2 n}{4\mu h^2}(1+o(1)).\end{aligned} \tag{A36}$$

Case 3: $h \ll \sqrt{\mu n}$ and $d < 2\mu(1-\mu/h)$. We optimize $c(h)$ here by again first optimizing over the distribution of h_i only, and then over the distribution of h_j. Therefore, up until (A30) we follow the same steps for optimizing over h_i. We then optimize for the distribution of h_j. Consider the majorant of the dual problem $\tilde{F}_1(x)$ described by $\lambda_1 = \frac{h^2(dn+2h\mu)}{4\mu^2 n^3}$, $\lambda_2 = \frac{h(h+2\mu)(dn+2h\mu)}{4\mu^2 n^3}$ and $\lambda_3 = -\frac{h^2(dn+2h\mu)}{4\mu n^3}$, giving as dual objective value

$$\lambda_1 d + \lambda_2 \mu + \lambda_3 = \frac{h\left(dh+2\mu^2\right)(dn+2h\mu)}{4\mu^2 n^3}. \tag{A37}$$

For the primal problem, consider the three-point distribution for h_2 of

$$\begin{aligned} p_0 &= \frac{d}{2\mu}, \quad p_\mu = 1 - \frac{d}{2\mu} - \frac{d}{2(h-\mu)}, \\ p_h &= \frac{d}{2(h-\mu)}, \end{aligned} \tag{A38}$$

which is a proper distribution as long as $d < 2\mu(1-\mu/h)$. Plugging this into (A30) for the distribution of h_j gives

$$\begin{aligned}&\mathbb{E}\left[\left(1-\frac{d}{2\mu}-\frac{d}{2(\mu n/h-\mu)}\right)\frac{h^2}{\mu n^3}h_2^2\right] + \mathbb{E}\left[\frac{d}{2(\mu n/h-\mu)}\frac{h_2^2}{\mu n}\right] \\ &= \frac{h\left(dh+2\mu^2\right)(dn+2h\mu)}{4\mu^2 n^3},\end{aligned} \tag{A39}$$

so that by strong duality, this is the optimal distribution for h_j. This yields for $c(h)$ that

$$\begin{aligned}\max_{\mathbb{P}_1,\mathbb{P}_2\in\mathcal{P}(\mu,d)} \mathbb{E}_{\mathbb{P}_1,\mathbb{P}_2}[c(h)] &= \frac{n^2}{h^2}\frac{h\left(dh+2\mu^2\right)(dn+2h\mu)}{4\mu^2 n^3} \\ &= \frac{d^2}{4\mu^2}(1+o(1)).\end{aligned} \tag{A40}$$

However, the maximal value of $c(h)$ is now attained by two different distributions for h_i and h_j, while our objective was to maximize $c(h)$ with i.i.d. distribution for h_i and h_j. We therefore now consider the uncorrelated three-point distribution $\mathbb{P}_3$ for h_i and h_j of

$$\begin{aligned} p_0 &= \frac{d}{2\mu}, \quad p_\mu = 1 - \frac{d}{2\mu} - \frac{d}{2(\sqrt{\mu n}-\mu)}, \\ p_{\sqrt{\mu n}} &= \frac{d}{2(\sqrt{\mu n}-\mu)}, \end{aligned} \tag{A41}$$

which is a proper distribution as long as $d < 2\mu(1-\sqrt{\mu/n})$. This yields as expected value for $c(h)$ of

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_3}[c(h)] &= \frac{n^2}{h^2}\frac{h^2}{(\mu n)^3}\mathbb{E}_{\mathbb{P}_3}[h^2]^2 \\ &= \frac{1}{\mu^3 n}\left(\frac{1}{2}d\sqrt{\mu n}+\mu^2\right)^2 \\ &= \frac{d^2}{4\mu^2}(1+o(1)),\end{aligned} \tag{A42}$$

which is the same leading order term as in (A40), where correlations between h_i and h_j are allowed. Since

$$\max_{\mathbb{P}_1,\mathbb{P}_2\in\mathcal{P}(\mu,d)} \mathbb{E}_{\mathbb{P}_1,\mathbb{P}_2}[c(h)] \geq \mathbb{E}_{\mathbb{P}_3}[c(h)], \tag{A43}$$

$\mathbb{P}_3$ is asymptotically optimal. Furthermore, an increase in d does not affect the set of feasible dual solutions for the unconstrained problem. Thus, (A37) is still an upper bound of the maximal $c(h)$. As by (A40) this dual value is asymptotically equal to $d^2/(4\mu^2)$, and (A42) achieves the same value asymptotically, this must imply that $\mathbb{P}_3$ is asymptotically optimal when it is a proper probability distribution, thus for all $d < 2\mu(1-\sqrt{\mu/n})$. □