Variational principle for scale-free network motifs

SCIENTIFIC REPORTS

OPEN

Variational principle for scale-free network motifs

Clara Stegehuis, Remco van der Hofstad & Johan S. H. van Leeuwaarden

Received: 20 September 2018
Accepted: 15 April 2019
Published online: 01 May 2019

For scale-free networks with degrees following a power law with an exponent $\tau \in (2, 3)$, the structures of motifs (small subgraphs) are not yet well understood. We introduce a method designed to identify the dominant structure of any given motif as the solution of an optimization problem. The unique optimizer describes the degrees of the vertices that together span the most likely motif, resulting in explicit asymptotic formulas for the motif count and its fluctuations. We then classify all motifs into two categories: motifs with small and large fluctuations.

Many real-world networks, like communication networks, social networks and biological networks, were found to be *scale free* with power-law degree distributions and infinite variance in the large-network limit[1–5]. The heavy tail of the power law comes with *hubs*, vertices of extremely high degree. Hubs create ultra-small distances, ultra-fast information spreading and resilience against random attacks, while the average node degree is small. Attested by real-world data and explained by mathematical models, the consequences of power-law connectivity patterns now belong to the foundations of network science.

Scale-free networks can be studied using random network models that connect vertices through edges forming power-law degree distributions. Connectivity patterns beyond edges are commonly described in terms of small subgraphs called motifs (or graphlets). There is increasing interest in algorithmic methods to count motifs[6–10], or the relation between motifs and network functions, like the spread of epidemics[11–16]. Motifs can describe the tendency for clustering and other forms of network organization[17–19].

Much existing work focuses on the classification of complex networks in terms of motif counts or frequencies. The occurrence of specific motifs such as wedges, triangles and cliques have been proven important for understanding real-world networks. Indeed, motif counts might vary considerably across different networks[20–22] and any given network has a set of statistically significant motifs. Statistical relevance can be expressed by comparing real-world networks to mathematically tractable models. A popular statistic takes the motif count, subtracts the expected number of motifs in the mathematical model, and divides by the standard deviation in the mathematical model[22–24]. Such a standardized test statistic predicts whether a motif is overrepresented in comparison to some mathematical model. This comparison, however, filters out the effect of the degree sequence, the network size and possibly other features that are needed to understand the structure and frequency of motifs.

With the goal to explain the occurrence of motifs beyond counting, we develop a method to identify, for any given motif, the composition that dominates the motif count as the solution of an optimization problem. The unique optimizer describes the degrees of the vertices that together span the most likely motif, as well as predicts the leading asymptotic order for the motif count in the large-network limit. Our method can potentially be applied to various random network models, but is developed first for the hidden-variable model[25–31], a random network model that generates graphs with power-law degrees. Given n vertices, the hidden-variable model associates to each node a hidden variable h drawn independently from the probability density

$$\rho(h) = Ch^{-\tau} \tag{1}$$

for some constant C and $h \geq h_{\min}$. Next, conditionally on all the hidden variables, each pair of vertices is joined independently with probability

$$p(h, h') = \min(hh'/(\mu n), 1). \tag{2}$$

with h and h' the hidden variables associated with the two vertices, and μ the mean of the hidden variables. For any given motif, we now seek for its most likely structure. The probability $P(H)$ of creating motif H on k uniformly chosen vertices can be written as

Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, Netherlands. Correspondence and requests for materials should be addressed to C.S. (email: c.stegehuis@tue.nl)

$$P(H) = \int_{\boldsymbol{h}} \mathbb{P}(H \text{ on } h_1, \ldots, h_k)\mathbb{P}(h_1, \ldots, h_k)\,\mathrm{d}\boldsymbol{h}, \tag{3}$$

where the integral is over all possible hidden-variable sequences on k vertices, with $\boldsymbol{h} = (h_1, \ldots, h_k)$ and $\mathbb{P}(h_1, \ldots, h_k)$ the density that a randomly chosen set of k hidden variables is proportional to $h_1, \ldots, h_k$. The degree of a node is asymptotically Poisson distributed with its hidden variable as mean[32], so (3) can be interpreted as a sum over all possible degree sequences. Therefore, our optimization method then needs to settle the following trade-off, inherently present in power-law networks: On the one hand, large-degree vertices contribute substantially to the number of motifs, because they are highly connected, and therefore participate in many motifs. On the other hand, large-degree vertices are by definition rare. This should be contrasted with lower-degree vertices that occur more frequently, but take part in fewer connections and hence fewer motifs. Therefore, our method give rise to a certain 'variational principle', because it finds the selection of vertices with specific degrees that together 'optimize' this trade-off and hence maximize the expected number of such motifs.

We leverage the optimization method in two ways. First, we derive sharp expressions for the motif counts in the large-network limit in terms of the network size and the power-law exponent. Second, we use the method to identify the fluctuations of motif counts.

We present two versions of the method that we call free and typical variation. Free variation corresponds to computing the average number of motifs over many samples of the random network model. Typical variation corresponds to the number of motifs in one single instance of the random graph model. Remarkably, for $\tau \in (2, 3)$ these can be rather different. After that, we apply the method to study motif count fluctuations. Finally, we provide a case study where we investigate the presence of motifs in some real-world network data.

Results

Free variation. We first show that only hidden-variable sequences $\boldsymbol{h}$ with hidden variables of specific orders give the largest contribution to (3). Write the hidden variables as $h_i \propto n^{\alpha_i}$ for some $\alpha_i \geq 0$ for all i. Then, using (2), the probability that motif H exists on vertices with hidden variables $\boldsymbol{h} = (n^{\alpha_1}, \ldots, n^{\alpha_k})$ satisfies

$$\mathbb{P}(H \text{ on } \boldsymbol{h}) \propto \prod_{(v_i,v_j)\in E_H:\alpha_i+\alpha_j<1} n^{\alpha_i+\alpha_j-1}. \tag{4}$$

The hidden variables are an i.i.d. sample from a power-law distribution, so that the probability that k uniformly chosen hidden variables satisfy $(h_1, \ldots, h_k) \propto (n^{\alpha_1}, \ldots, n^{\alpha_k})$ is of the order $n^{(1-\tau)\sum_i \alpha_i}$ (see Supplementary Material 2). Taking the product of this with (4) shows that the maximum contribution to the summand in (3) is obtained for those $\alpha_i \geq 0$ that maximize the exponent

$$(1-\tau)\sum_i \alpha_i + \sum_{(i,j)\in E_H:\alpha_i+\alpha_j<1} (\alpha_i + \alpha_j - 1), \tag{5}$$

which is a piecewise-linear function in α. In Supplementary Material 2, we show that the maximizer of this optimization problem satisfies $\alpha_i \in \left\{0, \frac{1}{2}, 1\right\}$ for all i. Thus, the maximal value of (5) is attained by partitioning the vertices of H into the sets S_1, S_2, S_3 such that vertices in S_1 have $\alpha_i = 0$, vertices in S_2 have $\alpha_i = 1$ and vertices in S_3 have $\alpha_i = \frac{1}{2}$. Then, the edges with $\alpha_i + \alpha_j < 1$ are edges inside S_1 and edges between S_1 and S_3. If we denote the number of edges inside S_1 by E_{S_1} and the number of edges between S_1 and S_3 by E_{S_1,S_3}, then maximizing (5) is equivalent to maximizing

$$B_f(H) = \max_{\mathcal{P}}\left[|S_1| - |S_2| - \frac{2E_{S_1} + E_{S_1,S_3}}{\tau - 1}\right] \tag{6}$$

over all partitions $\mathcal{P}$ of the vertices of H into S_1, S_2, S_3. This gives the following theorem (a more elaborate version is proven in Supplementary Material 2):

Theorem 1 (Expected motif count). *Let H be a motif on k vertices such that the solution to (6) is unique. As $n \to \infty$, the expected number of motifs H grows as*

$$\mathbb{E}[N(H)] = n^k P(H) \propto n^{\frac{3-\tau}{2}k + \frac{\tau-1}{2}B_f(H)}, \tag{7}$$

and is thus fully determined by the partition $\mathcal{P}^*$ that optimizes (6).

Theorem 1 implies that the expected number of motifs is dominated by motifs on vertices with hidden variables (and thus degrees) of specific orders of magnitude: constant degrees, degrees proportional to $\sqrt{n}$ or degrees proportional to n. Figures 1 and 2 show the partitions $\mathcal{P}^*$ that dominate the expected number of motifs on three, four and five vertices.

Typical variation. The largest degrees (hubs) in typical samples of the hidden-variable model scale as $n^{1/(\tau-1)}$ with high probability. The expected number of motifs, however, may be dominated by network samples where the largest degree is proportional to n (see Theorem 1). These samples contain many motifs because of the high degrees, and therefore contribute significantly to the expectation. Nevertheless, the probability of observing such a network tends to zero as n grows large. We therefore now adapt the variational principle with the goal to characterize the typical motif structure and hence the typical number of motifs.

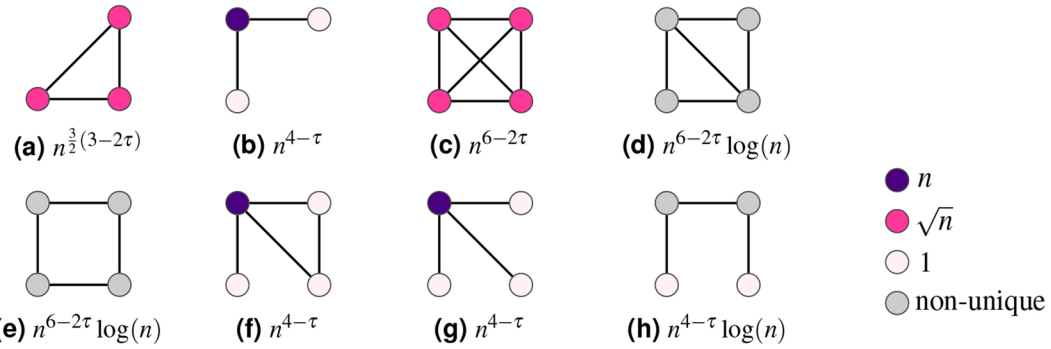

Figure 1. Scaling of the expected number of motifs on 3 and 4 vertices in n, where the vertex color indicates the dominant vertex degree. Vertices where the optimizer is not unique are gray.

We again assume degrees to be proportional to n^{α_i}, but now limit to degree sequences where the maximal degree is of order $n^{1/(\tau-1)}$, the natural cutoff in view of the typical hub degrees. The dominant typical motif structure is then obtained by maximizing (5), with the additional constraint that $\alpha_i \leq \frac{1}{\tau-1}$. In Supplementary Material 3 we show that the possible optimizers are one of four values $\alpha_i \in \left\{0, \frac{\tau-2}{\tau-1}, \frac{1}{2}, \frac{1}{\tau-1}\right\}$, and obtain an optimization problem similar to (6).

This shows that the typical degree of a motif is of constant order or proportional to $n^{1/(\tau-1)}$, $\sqrt{n}$ or $n^{(\tau-2)/(\tau-1)}$. Figure 3 and Supplementary Fig. 1 show the most likely motifs on three, four and five vertices. Observe that the dominant structures and the number of motifs of Figs 1 and 3 may differ. For example, the scaling of the expected number of claws (Fig. 1b) and the typical number of claws (Fig. 3b) is different. This is caused by the left upper vertex that has degree proportional to n in the free dominant structure, whereas its typical degree is proportional to $n^{1/(\tau-1)}$. Only when the solution to (6) does not involve hub vertices, the two scalings coincide. Hub vertices in the dominant structure give a major contribution to the motif count. While typical hub degrees scale as $n^{1/(\tau-1)}$, expected hub degrees may be much larger, causing the number of such motifs with hubs to scale faster in the free variation setting than in the typical variation setting. This indicates that the average and median motif count can differ dramatically.

Graphlets. It is also possible to only count the number of times H appears as an induced subgraph, also called graphlet counting. This means that an edge that is not present in graphlet H, should also be absent in the network motif. In Supplementary Material 4 we classify the expected and typical number of graphlets with a similar variational principle as for motifs. Supplementary Fig. 2 shows the typical behavior of graphlets on 4 vertices. This figure also shows that graphlet counting is more detailed than motif counting. For example, counting all square motifs is equivalent to counting all graphlets that contain the square as an induced subgraph: the square, the diamond and K_4. Indeed, we obtain that the number of square motifs scales as $n^{6-2\tau}\log(n)$ by adding the number of square, diamond and K_4 graphlets from Supplementary Fig. 2. This shows that most square motifs are actually the diamond graphlets of Supplementary Fig. 2b. Thus, graphlet counting gives more detailed information than motif counting.

Fluctuations. Self-averaging network properties have relative fluctuations that tend to zero as the network size n tends to infinity. Several physical quantities in for example Ising models, fluid models and properties of the galaxy display non-self-averaging behavior[33–39]. We consider motif counts $N(H)$ and call $N(H)$ self-averaging when $\mathrm{Var}(N(H))/\mathbb{E}[N(H)]^2 \to 0$ as $n \to \infty$. Essential understanding of $N(H)$ can then be obtained by taking a large network sample, since the sample-to-sample fluctuations vanish in the large-network limit. In contrast, if $\mathrm{Var}(N(H))/\mathbb{E}[N(H)]^2$ approaches a constant or tends to infinity as $n \to \infty$, the motif count is called non-self-averaging, in which case $N(H)$ shows (too) strong sample-to-sample fluctuations that cannot be mitigated by taking more network samples.

Our variational principle facilitates a systematic study of such fluctuations, and leads to a classification into self-averaging and non-self-averaging for all motifs H. It turns out that whether $N(H)$ is self-averaging or not depends on the power-law exponent τ and the dominant structure of H. We also show that non-self-averaging behavior of motif counts may not have the intuitive explanation described above. In some cases, motif counts in two instances are similar with high probability, but rare network samples behave differently, causing the motif count to be non-self-averaging. Thus, the classification of network motifs into self-averaging and non-self-averaging motifs does not give a complete picture of the motif count fluctuations. We therefore further divide the non-self-averaging motifs into two classes based on the type of fluctuations in the motif counts.

For a given motif H, let $H_1, \ldots, H_m$ denote all possible motifs that can be constructed by merging two copies of H at one or more vertices. We can then write the variance of the motif count as (see[37,40–42] and the Methods section)

$$\mathrm{Var}(N(H)) = C_1\mathbb{E}[N(H_1)] + \cdots + C_m\mathbb{E}[N(H_m)] + \mathbb{E}[N(H)]^2 O(n^{-1}). \tag{8}$$

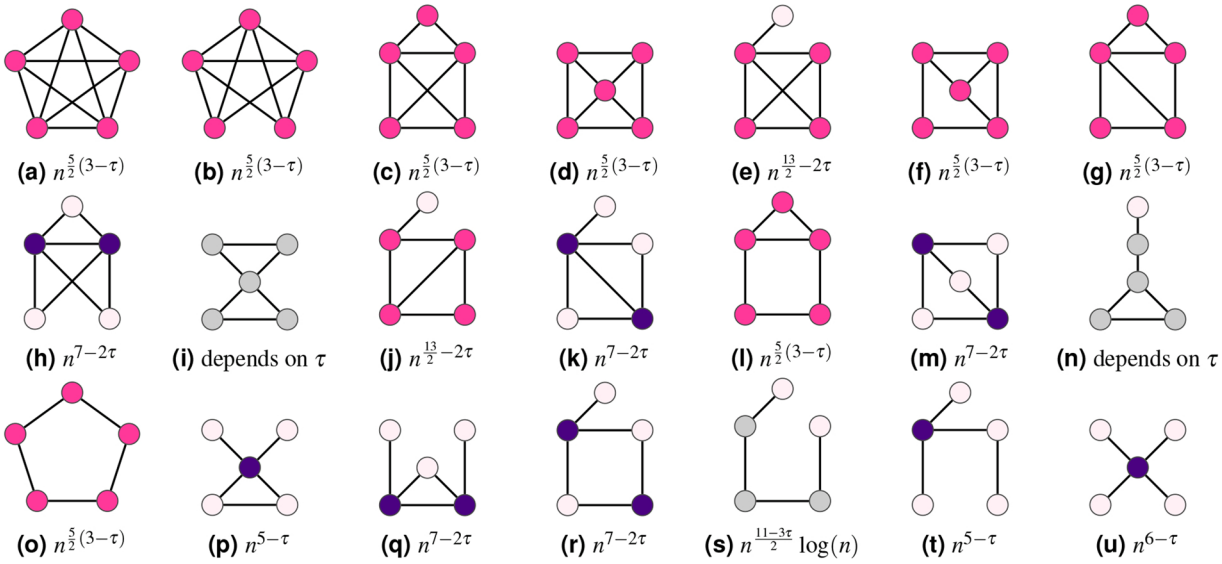

Figure 2. Scaling of the expected number of motifs on 5 vertices in n, where the vertex color indicates the dominant vertex degree, as in Fig. 1.

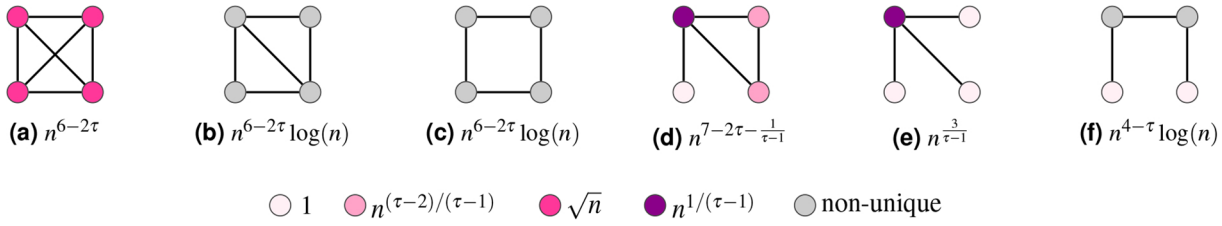

Figure 3. Typical scaling of the number of motifs on three or four vertices in n. The vertex color indicates the dominant vertex degree.

Self-averaging for	Subfigs of Fig. 1	Subfigs of Fig. 2
(2, 3)	c	a, b, c, d
(2, 5/2)	a	f, g, l, o
(2, 7/3)		i
—	b, d, e, f, g, h	e, h, j, k, m, n, p, q, r, s, t, u

Table 1. The values of $\tau \in (2, 3)$ where the motifs of Figs 1 and 2 are self-averaging.

for constants $C_1, \ldots, C_m$. Using (8), we can determine for any motif H whether it is self-averaging or not. First, we find all motifs that are created by merging two copies of H. For the triangle motif for example, these motifs are the bow-tie, where two triangles are merged at one single vertex, the diamond of Fig. 3b, and the triangle itself. We find the order of magnitude of the expected number of these motifs using Theorem 1 to obtain the variance of $N(H)$. We divide by $\mathbb{E}[N(H)]^2$, also obtained by Theorem 1, and check whether this fraction is diverging or not. Table 1 shows for which values of $\tau \in (2, 3)$ the motifs on 3, 4 and 5 vertices are self-averaging. For example, the triangle turns out to be self-averaging only for $\tau \in (2, 5/2)$.

Here is a general observation that underlines the importance of the dominant motif structure:

Theorem 2. *All self-averaging motifs for any $\tau \in (2, 3)$ have dominant free variation structures that consist of vertices with hidden variables $\Theta(\sqrt{n})$ only.*

We prove this theorem in Supplementary Material 5. Note that the condition on the dominant motif structure is a necessary condition for being self-averaging, but it is not a sufficient one, as the triangle example shows. Table 1 shows the values of τ for which all connected motifs on 3, 4 and 5 vertices are self-averaging. Combining the classification of the motifs into self-averaging and non-self-averaging with the classification based on the value of $B_f(H)$ from (6) as well as the difference between the expected and typical number of motifs yields a classification into the following three types of motifs:

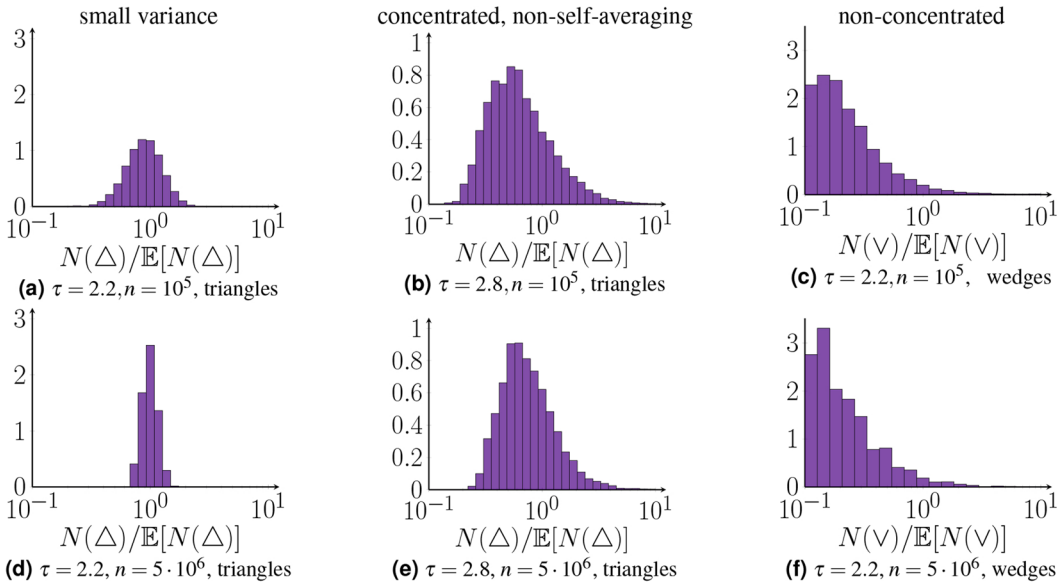

Figure 4. Density approximation of the normalized triangle and wedge counts for various values of τ and n, obtained over 10^4 network samples.

	n	m	τ
Gowalla	196591	950327	2.65
Oregon	11174	23409	2.08
Enron	36692	183831	1.97
PGP	10680	24316	2.24
Hep	9877	25998	3.50

Table 2. Statistics of the five data sets, where n is the number of vertices, m the number of edges, and τ the power-law exponent fitted by the procedure of[5].

Type I: Motifs with small variance. $B_f(H) = 0$ and $\mathrm{Var}(N(H))/\mathbb{E}[N(H)]^2 \to 0$. These motifs only contain vertices of degrees $\Theta(\sqrt{n})$. The number of such rescaled motifs converges to a constant[43]. Furthermore, the variance of the number of motifs is small compared to the second moment, so that the fluctuations of these types of motifs are quite small and vanish in the large network limit. The triangle for $\tau < 5/2$ is an example of such a motif, shown in Fig. 4b,e.

Type II: Concentrated, non-self-averaging motifs. $B_f(H) = 0$ and $\mathrm{Var}(N(H))/\mathbb{E}[N(H)]^2 \not\to 0$. These motifs also only contain vertices of degrees $\sqrt{n}$. Again, the rescaled number of such motifs converges to a constant in probability[43]. Thus, most network samples contain a similar amount of motifs as n grows large, even though these motifs are non-self-averaging. Still, in rare network samples the number of motifs significantly deviates from its typical number, causing the variance of the number of motifs to be large. Figure 4a,d illustrate this for triangle counts for $\tau \geq 5/2$. The fluctuations are larger than for the concentrated motifs, but most of the samples have motif counts close to the expected value.

Type III: Non-concentrated motifs. $B_f(H) > 0$. These motifs contain hub vertices. The expected and typical number of such motifs therefore scale differently in n. By Theorem 2, these motifs are non-self-averaging. The rescaled number of such motifs may not converge to a constant, so that two network samples contain significantly different motif counts. Figure 4c,f show that the fluctuations of these motifs are indeed of a different nature, since most network samples have motif counts that are far from the expected value.

Data. We now investigate motifs in five real-world networks with heavy-tailed degree distributions: the Gowalla social network[44], the Oregon autonomous systems network[44], the Enron email network[44,45], the PGP web of trust[46] and the High Energy Physics collaboration network (HEP)[44]. Table 2 provides detailed statistics of these data sets. Because the number of motifs can be obtained from the number of graphlets, we focus on graphlet counts. Figure 5 shows the graphlet counts on a logarithmic scale. The order of the graphlets is from the most occurring graphlet (the claw), to the least occurring graphlet (the square and K_4) in the hidden-variable model, see Supplementary Fig. 2. In three networks the motif ordering follows that of the hidden-variable model, while in two networks the ordering is different. In the HEP collaboration network, for example, K_4 occurs more frequently

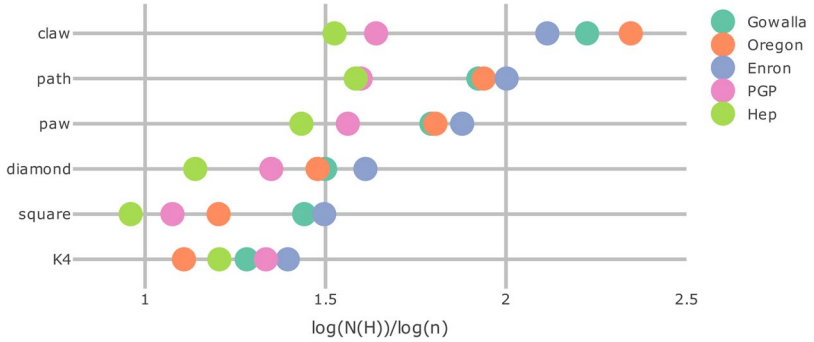

Figure 5. Number of graphlets on four vertices in five data sets on logarithmic scale: $\log(N(H))/\log(n)$. The ordering of the six different graphlets is from the most occurring in the hidden-variable model to the least. The values of the graphlet counts are presented in Supplementary Table 1.

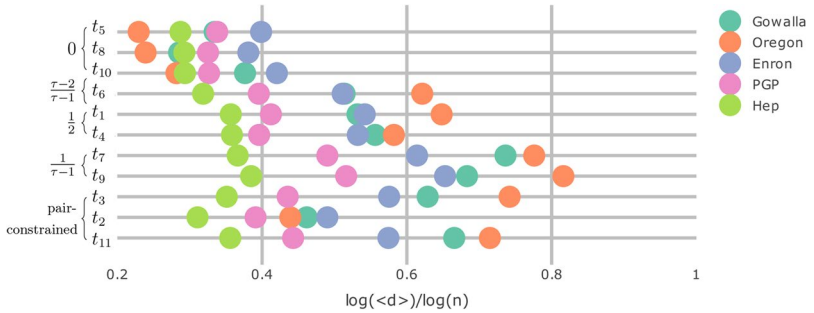

Figure 6. Average degree of the vertex types displayed in Supplementary Fig. 2 in 5 data sets on logarithmic scale. The curly brackets indicate the typical degree exponent of the vertex type in the hidden-variable model.

than the square. While this is not predicted by the hidden-variable model, it naturally arises due to the frequently occurring collaboration between four authors, which creates K_4 instead of the square. It would be interesting to see if this deviation from the ordering of the hidden-variable model can be linked to the specific nature of the data set in other examples.

Supplementary Fig. 2 enumerates all possible vertex types in graphlets on 4 vertices. In the hidden-variable model, vertex types t_7 and t_9 have typical degrees proportional to $n^{1/(\tau-1)}$, vertex types t_1 and t_4 typically have degrees proportional to $\sqrt{n}$, vertex type t_6 typically has degree proportional to $n^{(\tau-2)/(\tau-1)}$ and vertex types t_5, t_8, t_{10}, t_{11} typically have constant degree. vertex types t_2, t_3 and t_{11} do not have a unique optimizer. The degrees of these vertex types are pair-constrained (see the proof of Lemma 3). Figure 6 shows the typical degree of all 11 vertex types in the five real-world data sets. Vertices with typical degree 1 in the hidden-variable model have the lowest degree in the five data sets. Vertices that have typical degree $n^{1/(\tau-1)}$ in the hidden-variable model also have the highest degree among all vertex types in these five real-world data sets. Thus, typical degrees of vertices in a graphlet roughly follow the same ordering as in the hidden-variable model in these data sets.

The High Energy Physics collaboration network does not have a large distinction between the degrees of the different vertex types. This may be related to the fact that this network has less heavy-tailed degrees than the other networks (see Table 2).

Discussion

By developing a variational principle for the dominant degree composition of motifs in the hidden-variable model, we have identified the asymptotic growth of motif counts and their fluctuations for all motifs. This allowed us to determine for which values of the degree exponent $\tau \in (2, 3)$ the number of motifs is self-averaging. We further divide the non-self-averaging motifs into two classes with substantially different concentration properties.

Hub vertices in dominant motif structures cause wild degree fluctuations and non-self-averaging behavior, so that large differences between the average motif count and the motif count in one sample of the random network model arise. Non-self-averaging motifs without a hub vertex show milder fluctuations.

We expect that the variational principle can be extended to different random graph models, such as the hyperbolic random graph, the preferential attachment model and random intersection graphs. For example, for the hyperbolic random graph, the dominant structure of complete graphs is known to be $\sqrt{n}$ degrees[47] like in the hidden-variable model, but the dominant structures of other motifs are yet unknown.

In this paper, we presented a case study for motifs on 4 vertices in five scale-free network data sets. It would be interesting to perform larger network data experiments to investigate whether motifs in real-world

network data also have typical vertex degrees, and to what extent these vertex degrees are similar to the ones of the hidden-variable model. Similarly, investigating the typical behavior of motifs on more than 4 vertices in real-world data compared to the hidden-variable model is another topic for future research.

It would also be interesting to develop statistical tests for the presence of motifs in real-world data using the results from this paper. For example, one could compare the ordering of all motifs for size k from the most frequent occurring to the least frequent occurring motif, and compare this to the ordering in a hidden-variable with the same degree-exponent. This could shed some light on which motifs in a given data set appear more often than expected.

Methods

Fluctuations. *Triangle fluctuations.* We first illustrate how we can apply the variational principle to obtain the variance of the number of subgraphs by computing the variance of the number of triangles in the hidden-variable model. Let Δ denote the number of triangles, and let $\Delta_{i,j,k}$ denote the event that vertices i, j and k form a triangle. Then, we can write the number of triangles as

$$\Delta = \frac{1}{6}\sum_{i,j,k\in[n]}{}' \Delta_{i,j,k}, \tag{9}$$

where $\sum'$ denotes the sum over distinct indices. Thus, the variance of the number of triangles can be written as

$$\mathrm{Var}(\Delta) = \sum_{i,j,k\in[n]}{}' \sum_{s,t,u\in[n]}{}' \mathbb{P}(\Delta_{i,j,k}, \Delta_{s,t,u}) - \mathbb{P}(\Delta_{i,j,k})\mathbb{P}(\Delta_{s,t,u}). \tag{10}$$

When i, j, k and s, t, u do not overlap, the hidden variables of i, j, k and s, t, u are independent, so that the event that i, j and k form a triangle and the event that s, t and u form a triangle are independent. Thus, when i, j, k, s, t, u are all distinct, $\mathbb{P}(\Delta_{i,j,k}, \Delta_{s,t,u}) = \mathbb{P}(\Delta_{i,j,k})\mathbb{P}(\Delta_{s,t,u})$, so that the contribution from 6 distinct indices to (10) is zero. On the other hand, when $i = u$ for example, the first term in (10) denotes the probability that a bow tie (see Fig. 2i) is present with i as middle vertex. Furthermore, since the degrees are i.i.d. and the edge statuses are independent as well, $\mathbb{P}(\Delta_{i,j,k})$ is the same for any $i \neq j \neq k$, so that

$$\mathbb{P}(\Delta_{i,j,k}) = \frac{\mathbb{E}[\Delta]}{6\binom{n}{3}} = \frac{\mathbb{E}[\Delta]}{6n^3}(1 + o(1)). \tag{11}$$

This results in

$$\begin{aligned}\mathrm{Var}(\Delta) &= 9\mathbb{E}[\#\text{ bow-ties}] - 9n^{-1}\mathbb{E}[\Delta]^2 + 18\mathbb{E}[\#\text{ diamonds}] \\ &\quad - 18n^{-2}\mathbb{E}[\Delta]^2 + 6\mathbb{E}[\Delta] - 6n^{-3}\mathbb{E}[\Delta]^2 \\ &= 9\mathbb{E}[\#\text{ bow-ties}] + 18\mathbb{E}[\#\text{ diamonds}] \\ &\quad + 6\mathbb{E}[\Delta] + \mathbb{E}[\Delta]^2 O(n^{-1}),\end{aligned} \tag{12}$$

where the diamond motif is as in Fig. 3b. The combinatorial factors 9, 18 and 6 arise because there are 9 ways to construct a bow tie (18 for a diamond, and 6 for a triangle) by letting two triangles overlap. The diamond motif does not satisfy the assumption in Theorem 1 that the optimal solution to (6) is unique. However, we can show the following result:

Lemma 3. $\mathbb{E}[\text{number of diamonds}] = \Theta(n^{6-2\tau})\log(n)$.

Proof. Let i and j be the vertices at the diagonal of the diamond, and k and s the corner vertices. Then (5) is optimized for $\alpha_i = \beta$, $\alpha_j = 1 - \beta$, $\alpha_k = \beta$ and $\alpha_s = 1 - \beta$ for all values of $\beta \in [1/2, 1]$ (see Supplementary Information 6). All these optimizers together give the major contribution to the number of diamonds. Thus, we need to find the number of sets of four vertices, satisfying

$$h_i h_j = \Theta(n), \quad h_i > h_j, \quad h_k = \Theta(h_i), \quad h_s = \Theta(h_j). \tag{13}$$

Given h_i and h_j, the number of sets of two vertices k, s with $h_k = \Theta(h_i)$ and $h_s = \Theta(h_j)$ is given by $n^2 h_i^{1-\tau} h_j^{1-\tau} = \Theta(n^{3-\tau})$, where we used that $h_i h_j = \Theta(n)$. The number of sets of vertices i, j such that $h_i h_j = \Theta(n)$ can be found using that the product of two independent power-law random variables is again distributed as a power law, with an additional logarithmic term[48], Eq. (2.16) (where in our setting equality holds in[48], Eq. (2.16), since we assume a pure power-law distribution). Thus, the number of sets of vertices with $h_i h_j = \Theta(n)$ scales as $n^2 n^{1-\tau}\log(n)$. Then, the expected number of sets of four vertices satisfying all constraints on the degrees scales as $n^{6-2\tau}\log(n)$. By (4), the probability that a diamond exists on degrees satisfying (13) is asymptotically constant, so that the expected number of diamonds also scales as $n^{6-2\tau}\log(n)$. □

Theorem 1 gives for the number of bow ties that (using[49] to find the optimal partition)

$$\mathbb{E}[\#\text{ bow ties}] = \begin{cases}\Theta(n^{\frac{5}{2}(3-\tau)}) & \tau < \frac{7}{3}, \\ \Theta(n^{4-\tau}) & \tau \geq \frac{7}{3},\end{cases} \tag{14}$$

and for the number of triangles (again using[49]) that $\mathbb{E}[\Delta] = \Theta(n^{3(3-\tau)/2})$. Combining this with (12) results in

$$\mathrm{Var}(\Delta) = \begin{cases} \Theta(n^{\frac{5}{2}(3-\tau)}) & \tau < \frac{7}{3}, \\ \Theta(n^{4-\tau}) & \tau \geq \frac{7}{3}. \end{cases} \tag{15}$$

To investigate whether the triangle motif is self-averaging, we need to compare the variance to the second moment of the number of triangles, which results in

$$\frac{\mathrm{Var}(\Delta)}{\mathbb{E}[\Delta]^2} = \begin{cases} \Theta(n^{\frac{1}{2}(\tau-3)}), & \tau < \frac{7}{3}, \\ \Theta(n^{2\tau-5}), & \tau \geq \frac{7}{3}. \end{cases} \tag{16}$$

Therefore,

$$\lim_{n\to\infty} \frac{\mathrm{Var}(\Delta)}{\mathbb{E}[\Delta]^2} = \begin{cases} 0 & \tau < \frac{5}{2}, \\ \infty & \tau > \frac{5}{2}. \end{cases} \tag{17}$$

For $\tau = 5/2$, the limit in (17) is of constant order of magnitude. Thus, the number of triangles is self-averaging as long as $\tau < \frac{5}{2}$. When $\tau \geq \frac{5}{2}$ the number of triangles is not self-averaging.

General motif fluctuations. We now compute the variance of general motifs, similar to the triangle example. Let $\boldsymbol{i} = (i_1, \ldots, i_k)$ be such that $i_p \neq i_q$ when $p \neq q$. We can then write the variance as

$$\begin{aligned} \mathrm{Var}(N(H)) = & \sum_{\boldsymbol{i}\in[n]^k} \sum_{\boldsymbol{j}\in[n]^k} (\mathbb{P}(H_{\boldsymbol{i}}, H_{\boldsymbol{j}} \text{present}) \\ & - \mathbb{P}(H_{\boldsymbol{i}} \text{present})\mathbb{P}(H_{\boldsymbol{j}} \text{present})). \end{aligned} \tag{18}$$

The sum splits into several cases, depending on the overlap of $\boldsymbol{i}$ and $\boldsymbol{j}$. The term where $\boldsymbol{i}$ and $\boldsymbol{j}$ do not overlap equals zero, since edges between vertices that do not overlap are independent.

Now suppose $\boldsymbol{i}$ and $\boldsymbol{j}$ overlap at $i_{t_1}, \ldots i_{t_r}$ and $j_{s_1}, \ldots, j_{s_r}$ for some $r > 0$. Then $\mathbb{P}(H_{\boldsymbol{i}}, H_{\boldsymbol{j}} \text{present})$ is equal to the probability that motif $\widetilde{H}$ is present on vertices $i_1, \ldots, i_k, j_1, \ldots, j_k \backslash j_{s_1}, \ldots, j_{s_r}$, where $\widetilde{H}$ denotes the motif that is constructed by merging two copies of H at i_{t_1} with j_{s_1}, at i_{t_2} with j_{s_2} and so on. Thus, this term can be written as

$$\sum_{t_1,\ldots,t_{2k-r}}{}' \mathbb{P}(\widetilde{H}_{t_1,\ldots,t_{2k-r}} \text{present}) = \mathbb{E}[N(\widetilde{H})], \tag{19}$$

where $\sum'$ denotes a sum over distinct indices. Furthermore, since the degrees are i.i.d. as well as the connection probabilities, $\mathbb{P}(H_i \text{present}) = \mathbb{E}[N(H)]/\binom{n}{k}$. Thus,

$$\sum_{t_1,\ldots,t_{2k-r}}{}' \mathbb{P}(H_{t_1,\ldots,t_k} \text{present})\mathbb{P}(H_{t_{k-r},\ldots,t_{2k-r}} \text{present}) = n^{-r}\mathbb{E}[N(H)]^2 O(1). \tag{20}$$

Let $H_1, \ldots, H_l$ denote all motifs that can be constructed by merging two copies of H at at least one vertex. We can then write the variance of the motif count as (see[37,40–42])

$$\mathrm{Var}(N(H)) = C_1\mathbb{E}[N(H_1)] + \cdots + C_l\mathbb{E}[N(H_l)] + \mathbb{E}[N(H)]^2 O(n^{-1}). \tag{21}$$

where C_i is a combinatorial constant that denotes the number of distinct ways to merge two copies of H into H_i. These constants satisfy[40]

$$\sum_{i=1}^{l} C_i = \sum_{s=0}^{k-1} \binom{k}{s}^2 (k-s)!. \tag{22}$$

References

1. Albert, R., Jeong, H. & Barabási, A.-L. Internet: Diameter of the world-wide web. *Nature* **401**, 130–131, https://doi.org/10.1038/43601 (1999).
2. Faloutsos, M., Faloutsos, P. & Faloutsos, C. On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, vol. 29, 251–262 (ACM, 1999).
3. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
4. Vázquez, A., Pastor-Satorras, R. & Vespignani, A. Large-scale topological and dynamical properties of the internet. *Phys. Rev. E* **65**, 066130, https://doi.org/10.1103/PhysRevE.65.066130 (2002).
5. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703, https://doi.org/10.1137/070710111 (2009).
6. Kashtan, N., Itzkovitz, S., Milo, R. & Alon, U. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* **20**, 1746–1758, https://doi.org/10.1093/bioinformatics/bth163 (2004).

7. Grochow, J. A. & Kellis, M. Network motif discovery using subgraph enumeration and symmetry-breaking. In *In RECOMB*, 92–106 (2007).
8. Omidi, S., Schreiber, F. & Masoudi-Nejad, A. MODA: An efficient algorithm for network motif discovery in biological networks. *Genes & Genetic Systems* **84**, 385–395, https://doi.org/10.1266/ggs.84.385 (2009).
9. Schreiber, F. & Schwobbermeyer, H. MAVisto: a tool for the exploration of network motifs. *Bioinformatics* **21**, 3572–3574, https://doi.org/10.1093/bioinformatics/bti556 (2005).
10. Wernicke, S. & Rasche, F. Fanmod: a tool for fast network motif detection. *Bioinformatics* **22**, 1152–1153, https://doi.org/10.1093/bioinformatics/btl038 (2006).
11. House, T., Davies, G., Danon, L. & Keeling, M. J. A motif-based approach to network epidemics. *Bull. Math. Biol.* **71**, 1693–1706, https://doi.org/10.1007/s11538-009-9420-z (2009).
12. Noël, P.-A., Allard, A., Hébert-Dufresne, L., Marceau, V. & Dubé, L. J. Spreading dynamics on complex networks: a general stochastic approach. *J. Math. Biol.* **69**, 1627–1660, https://doi.org/10.1007/s00285-013-0744-9 (2013).
13. Zhang, J. & Moura, J. M. Subgraph density and epidemics over networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, https://doi.org/10.1109/icassp.2014.6853772 (IEEE, 2014).
14. Ritchie, M., Berthouze, L. & Kiss, I. Z. Beyond clustering: mean-field dynamics on networks with arbitrary subgraph composition. *J. Math. Biol.* **72**, 255–281, https://doi.org/10.1007/s00285-015-0884-1 (2015).
15. Ritchie, M., Berthouze, L. & Kiss, I. Z. Generation and analysis of networks with a prescribed degree sequence and subgraph family: higher-order structure matters. *J. Complex Networks* cnw011, https://doi.org/10.1093/comnet/cnw011 (2016).
16. Stegehuis, C., van der Hofstad, R. & van Leeuwaarden, J. S. H. Epidemic spreading on complex networks with community structures. *Sci. Rep.* **6**, 29748, https://doi.org/10.1038/srep29748 (2016).
17. Ravasz, E. & Barabási, A.-L. Hierarchical organization in complex networks. *Phys. Rev. E* **67**, 026112, https://doi.org/10.1103/PhysRevE.67.026112 (2003).
18. Benson, A. R., Gleich, D. F. & Leskovec, J. Higher-order organization of complex networks. *Science* **353**, 163–166 (2016).
19. Tsourakakis, C. E., Pachocki, J. & Mitzenmacher, M. Scalable motif-aware graph clustering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 1451–1460, https://doi.org/10.1145/3038912.3052653 (2017).
20. Milo, R. *et al*. Network motifs: Simple building blocks of complex networks. *Science* **298**, 824–827, https://doi.org/10.1126/science.298.5594.824 (2002).
21. Wuchty, S., Oltvai, Z. N. & Barabási, A.-L. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat. Genet.* **35**, 176–179, https://doi.org/10.1038/ng1242 (2003).
22. Milo, R. *et al*. Superfamilies of evolved and designed networks. *Science* **303**, 1538–1542, https://doi.org/10.1126/science.1089167 (2004).
23. Onnela, J.-P., Saramäki, J., Kertész, J. & Kaski, K. Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E* **71**, 065103, https://doi.org/10.1103/PhysRevE.71.065103 (2005).
24. Gao, C. & Lafferty, J. Testing network structure using relations between small subgraph probabilities. *arXiv:1704.06742* (2017).
25. Goh, K.-I., Kahng, B. & Kim, D. Universal behavior of load distribution in scale-free networks. *Physical Review Letters* **87**, https://doi.org/10.1103/physrevlett.87.278701 (2001).
26. Boguñá, M. & Pastor-Satorras, R. Class of correlated random networks with hidden variables. *Phys. Rev. E* **68**, 036112, https://doi.org/10.1103/PhysRevE.68.036112 (2003).
27. Park, J. & Newman, M. E. J. Statistical mechanics of networks. *Phys. Rev. E* **70**, 066117, https://doi.org/10.1103/PhysRevE.70.066117 (2004).
28. Norros, I. & Reittu, H. On a conditionally Poissonian graph process. *Adv. Appl. Probab.* **38**, 59–75, https://doi.org/10.1017/s000186780000080x (2006).
29. Chung, F. & Lu, L. The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci. USA* **99**, 15879–15882 (2002).
30. Söderberg, B. General formalism for inhomogeneous random graphs. *Physical Review E* **66**, https://doi.org/10.1103/physreve.66.066121 (2002).
31. Caldarelli, G., Capocci, A., Rios, P. D. L. & Muñoz, M. A. Scale-free networks from varying vertex intrinsic fitness. *Physical Review Letters* **89**, https://doi.org/10.1103/physrevlett.89.258702 (2002).
32. Stegehuis, C., van der Hofstad, R., van Leeuwaarden, J. S. H. & Janssen, A. J. E. M. Clustering spectrum of scale-free networks. *Phys. Rev. E* **96**, 042309, https://doi.org/10.1103/physreve.96.042309 (2017).
33. Wiseman, S. & Domany, E. Lack of self-averaging in critical disordered systems. *Phys. Rev. E* **52**, 3469–3484, https://doi.org/10.1103/physreve.52.3469 (1995).
34. Wiseman, S. & Domany, E. Finite-size scaling and lack of self-averaging in critical disordered systems. *Phys. Rev. Lett.* **81**, 22–25, https://doi.org/10.1103/physrevlett.81.22 (1998).
35. Aharony, A. & Harris, A. B. Absence of self-averaging and universal fluctuations in random systems near critical points. *Phys. Rev. Lett.* **77**, 3700–3703, https://doi.org/10.1103/physrevlett.77.3700 (1996).
36. Das, M. & Green, J. R. Self-averaging fluctuations in the chaoticity of simple fluids. *Phys. Rev. Lett.* **119**, https://doi.org/10.1103/physrevlett.119.115502 (2017).
37. Ostilli, M. Fluctuation analysis in complex networks modeled by hidden-variable models: Necessity of a large cutoff in hidden-variable models. *Phys. Rev. E* **89**, 022807, https://doi.org/10.1103/PhysRevE.89.022807 (2014).
38. Pastur, L. A. & Shcherbina, M. V. Absence of self-averaging of the order parameter in the Sherrington-Kirkpatrick model. *J. Statist. Phys.* **62**, 1–19, https://doi.org/10.1007/bf01020856 (1991).
39. Labini, F. S., Vasilyev, N. L., Pietronero, L. & Baryshev, Y. V. Absence of self-averaging and of homogeneity in the large-scale galaxy distribution. *EPL Europhysics Lett.* **86**, 49001, https://doi.org/10.1209/0295-5075/86/49001 (2009).
40. Frank, O. Moment properties of subgraph counts in stochastic graphs. *Ann. New York Acad. Sci.* **319**, 207–218, https://doi.org/10.1111/j.1749-6632.1979.tb32791.x (1979).
41. Picard, F., Daudin, J.-J., Koskas, M., Schbath, S. & Robin, S. Assessing the exceptionality of network motifs. *J. Comput. Biol.* **15**, 1–20 (2008).
42. Matias, C., Schbath, S., Birmelé, E., Daudin, J.-J. & Robin, S. Network motifs: mean and variance for the count. *Revstat* **4**, 31–51 (2006).
43. van der Hofstad, R., van Leeuwaarden, J. S. H. & Stegehuis, C. Optimal subgraph structures in scale-free configuration models. *arXiv:1709.03466* (2017).
44. Leskovec, J. & Krevl, A. SNAP Datasets: Stanford large network dataset collection, http://snap.stanford.edu/data, Date of access: 14/03/2017 (2014).
45. Klimt, B. & Yang, Y. Introducing the Enron Corpus. In *CEAS* (2004).
46. Boguñá, M., Pastor-Satorras, R., Daz-Guilera, A. & Arenas, A. Models of social networks based on social distance attachment. *Phys. Rev. E* **70**, https://doi.org/10.1103/physreve.70.056122 (2004).
47. Friedrich, T. & Krohmer, A. Cliques in hyperbolic random graphs. In *INFOCOM proceedings 2015*, 1544–1552 (IEEE, 2015).
48. van der Hofstad, R. & Litvak, N. Degree-degree dependencies in random graphs with heavy-tailed degrees. *Internet Math.* **10**, 287–334, https://doi.org/10.1080/15427951.2013.850455 (2014).
49. Stegehuis, C. Dominant motif structures, code for solving optimization problem (6), https://doi.org/10.5281/zenodo.2545464 (2019).

Acknowledgements
This work is supported by NWO TOP grant 613.001.451 and by the NWO Gravitation Networks grant 024.002.003. The work of R.v.d.H. is further supported by the NWO VICI grant 639.033.806. The work of J.v.L. is further supported by an NWO TOP-GO grant and by an ERC Starting Grant.

Author Contributions
C.S. performed all numerical simulations. C.S., R.v.d.H. and J.v.L. wrote the manuscript.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-43050-8.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.