Random Overlapping Communities: Approximating Motif Densities of Large Graphs

Samantha Petti∗  Santosh Vempala †

September 28, 2017

Abstract

A wide variety of complex networks (social, biological, information etc.) exhibit local clustering with substantial variation in the clustering coefficient (the probability of neighbors being connected). Existing models of large graphs capture power law degree distributions (Barabási-Albert) and small-world properties (Watts-Strogatz), but only limited clustering behavior. We introduce a generalization of the classical Erdős-Rényi model of random graphs which provably achieves a wide range of desired clustering coefficient, triangle-to-edge and four-cycle-to-edge ratios for any given graph size and edge density. Rather than choosing edges independently at random, in the Random Overlapping Communities model, a graph is generated by choosing a set of random, relatively dense subgraphs (“communities”). We discuss the explanatory power of the model and some of its consequences.

1 Introduction

Randomness has been an effective metaphor to model and understand the structure of complex networks. In 1959, Erdős and Rényi [9, 10] defined the simple random graph model $G_{n,p}$, where every pair of $n$ vertices is independently connected with probability $p$. Their seminal work transformed the field of combinatorics and laid the foundation of network science. Mathematicians have extensively studied properties of graphs generated from this model and used it to prove the existence of graphs with certain properties. (See [11] for a survey.) The comparison of real-world graphs to $G_{n,p}$ is a popular tool for highlighting their nonrandom features [27, 19, 20]. Moreover, the model has inspired more sophisticated random graph models, as predicted by Erdős and Rényi in the following remark from their pre-internet/pre-social graphs article:

This may be interesting not only from a purely mathematical point of view ... if one aims at describing such a real situation, one should replace the hypothesis of equiprobability of all connections by some more realistic hypothesis. It seems plausible that by considering the random growth of more complicated structures one could obtain fairly reasonable models of more complex real growth processes.

The two most influential random graph models designed to mimic properties of real-world graphs are the Watts-Strogatz small world model [28] and the Barabási-Albert preferential attachment model.

∗Georgia Tech, spetti@gatech.edu. Supported in part by an NSF graduate fellowship.
†Georgia Tech, vempala@gatech.edu. Both authors were supported in part by NSF awards CCF-1563838 and CCF-1717349.
model \[7\]. Briefly, the first is a process that randomly rewires connections of a regular ring lattice graph. The resulting graphs have small diameter and high clustering coefficient (the probability that two neighbors of a randomly selected vertex are adjacent). The second is a growth model that repeatedly adds a new vertex to an existing graph and connects to existing vertices with probability proportional to their degree. This model exhibits and maintains a power law in the distribution of vertex degrees, another commonly observed phenomenon.

These and other existing random graph models do not capture the following fundamental aspects of local structure: (1) Existing models cannot be tuned to produce graphs with arbitrary density, triangle-to-edge ratio, and four-cycle-to-edge ratio. (2) The clustering coefficients of graphs produced by existing models lie in very limited ranges determined by the graph’s density. In reality, the clustering coefficients of a variety of complex graphs (social, biological, information etc.) vary substantially and are not simply a function of the graph’s density \[19\].

We introduce the Random Overlapping Communities (ROC) model, a simple generalization of the Erdős-Rényi model, which produces graphs with a wide range of clustering coefficients as well as triangle-to-edge and four-cycle-to-edge ratios. The model generates graphs that are the union of many relatively dense random communities. A community is an instance of \(G_{s,q}\) on a set of \(s\) randomly chosen vertices. A ROC graph is the union of many randomly selected communities that overlap, so each vertex is a member of multiple communities. The size \(s\) and density \(q\) of the communities determine clustering coefficient and triangle and four-cycle ratios.

Capturing motif densities. A widely-used technique for inferring the structure and function of a graph is to observe overrepresented motifs, i.e., small patterns (subgraphs) that appear frequently. Recent work describes the overrepresented motifs of a variety of graphs including transcription regulation graphs, protein-protein interaction graphs, the rat visual cortex, ecological food webs, and the internet (WWW), \[30, 5, 25, 17\]. The type of overrepresented motifs has been shown to be correlated with the graph’s function \[17\]. A model that produces graphs with high motif counts is necessary for approximating graphs whose function depends on the abundance of a particular motif. Here we focus on the two most basic motifs—triangles and four-cycles.

A natural approach to constructing a graph with high motif density is to repeatedly add the motif on a randomly chosen subset of vertices. However, this process yields low motif to edge ratios for sparse graphs. For example, a graph on \(n\) vertices with average degree less than \(\sqrt{n}\) built by randomly adding triangles will have a triangle-to-edge-ratio at most \(2/3\). (See Theorem \[12\].) In \[21\] Newman considers a similar approach which produces graphs with varied degree sequences and triangle to edge ratio strictly less than \(1/3\). However, it is not hard to construct graphs with arbitrarily high triangle ratio (growing with the size of the graph).

In the dense setting, a constant-size stochastic block model can be used to approximate graphs with high motif densities, as guaranteed by Szemerédi’s regularity lemma (see \[15\]). In a stochastic block model \(M\), each vertex is assigned to one of \(k\) classes, and an edge is added between each pair of vertices independently with probability \(M_{i,j}\) where \(i\) and \(j\) are the classes of the vertices. However, the situation is drastically different for nondense graphs. To construct a sparse graph with maximum degree at most \(n^{1/3}\) with non-vanishing four-cycle density, the rank of \(M\) must grow with the size of the graph.

**Theorem 1.** Let \(M\) be a symmetric \(n \times n\) matrix with entries in \([0, 1]\) such that each row sum is at most \(d\). Let \(G\) be a graph on \(n\) vertices obtained by adding each edge \((i, j)\) independently with probability \(M_{ij}\). Then the expected number of \(k\)-cycles in \(G\) at most \(d^{\text{rank}(M)}\).
Figure 1: The clustering coefficient in real world graphs is much greater than that of an E-R random graph of the same density. Data from Table 3.1 of [19].

For example, the $d$-dimensional hypercube graph on $n = 2^d$ vertices has a $\log(n)/4$ four-cycle-to-edge ratio; a stochastic block model $M$ that produces a graph of the same size, degree, and ratio must have rank at least $O(n/\log^2 n)$.

In contrast to the above approaches, the ROC model produces graphs with arbitrary triangle and four-cycle ratios independent of the density or size of the graph. In Theorem 3 we show that for almost all triangle and four-cycle ratios arising from some graph, there exists parameters for the ROC model to produce graphs with these ratios, simultaneously. Moreover, the vanishing set of triangle and four-cycle ratio pairs not achievable exactly can be approximated to within a small error.

**Clustering coefficient.** The clustering coefficient at a vertex $v$ is the probability two randomly selected neighbors are adjacent:

$$C(v) = \frac{|\{(a, b) : a, b \in N(v), a \sim b\}|}{\deg(v)(\deg(v) - 1)/2}.$$  

Equivalently the clustering coefficient is twice the ratio of the number of triangles containing $v$ to the degree of $v$ squared. The ROC model is well suited to produce random graphs that reflect the high average clustering coefficients of real world graphs. Figure 1 illustrates the markedly high clustering coefficients of real-world graphs as compared with Erdős-Rényi (E-R) graphs of the same density. In Theorem 4 we prove the average clustering coefficient of a ROC graph is approximately $sq^2/d$, meaning that tuning the parameters $s$ and $q$ with $d$ fixed yields wide range of clustering coefficients for a fixed density. Furthermore, Theorem 5 describes the inverse relationship between degree and clustering coefficient in ROC graphs, a phenomena observed in protein-protein interaction graphs, the internet, and various social graphs [26, 16, 18, 2].

**Structure of the paper.** In Section 2 we introduce the ROC model, and then in Section 3 we show the model’s ability to produce graphs with specified size, density, triangle and four-cycle ratios and clustering coefficients. In Section 4 we introduce a variation of the ROC model which
produces graphs with various degree distributions and tunable clustering coefficient. We end with a discussion of the model’s mathematical interest and explanatory value in real-world settings in Section 5.

2 The Random Overlapping Communities model

A complex graph is modeled as the union of relatively dense, random communities. More precisely, to construct a graph on \( n \) vertices with expected degree \( d \), we pick \( dn/(qs(s - 1)) \) random graphs, each of density \( q \) on a random subset of \( s \) of the \( n \) vertices.

\[
\text{ROC}(n, d, s, q).
\]

Output: a graph on \( n \) vertices with expected degree \( d \).

Repeat \( dn/(qs(s - 1)) \) times:

1. Pick a random subset \( S \) of vertices (from \( \{1, 2, \ldots, n\} \)) by selecting each vertex with probability \( s/n \).
2. Add the random graph \( G_{|S|,q} \) on \( S \), i.e., for each pair in \( S \), add the edge between them independently with probability \( q \); if the edge already exists, do nothing.

This generalizes the standard E-R model, which is the special case when \( s = n \) and a single community is picked. For \( G \sim \text{ROC}(n, d, s, q) \) the expected degree of each vertex is \( d \). If \( d > sq \log n \) then with high probability \( G \) will be connected. Moreover if \( d/p > \log \frac{nd}{s(s-1)p} \), then with high probability the communities of \( G \) will be connected even though there may be isolated vertices.
3 Approximation by ROC graphs

In this section we analyze small cycle counts and local clustering coefficient of ROC graphs. For proofs of the theorems refer to Section C of the appendix. We state our results as they hold asymptotically with respect to \( n \).

3.1 Triangle and four-cycle count in ROC graphs.

Define \( R_k \) as the ratio between the number of \( k \) cycles and the edges in a graph:

\[
R_k(G) = \frac{C_k(G)}{|E(G)|},
\]

where \( C_k(G) \) denotes the number of \( k \) cycles in \( G \). For \( G \sim ROC(n,d,s,q) \), we instead define

\[
\overline{R}_k(G) = \frac{2E[C_k(G)]}{nd},
\]

the ratio of the expected number of \( k \) cycles to the expected number of edges.

Lemma 2. Let \( G \sim ROC(n,d,s,q) \) and \( s = \omega(1) \). Then

\[
\lim_{n \to \infty} \overline{R}_3(G) = \frac{sq^2}{3} \text{ for } d = o(\sqrt{n}) \quad \text{and} \quad \lim_{n \to \infty} \overline{R}_4(G) = \frac{s^2q^3}{4} \text{ for } d = o(n^{1/3}).
\]

By varying \( s \) and \( q \), we can construct a ROC graph that achieves any ratio of triangles to edges or any ratio of four-cycles to edges. By setting \( s = \sqrt{\log(n)/4} \) and \( q = 1 \), we obtain a family of graphs with the hypercube four-cycle-to-edge ratio \( \log(n)/4 \), something not possible with any existing random graph model.

Moreover, it is possible to achieve a given ratio by larger, sparser communities or by smaller, denser communities. For example communities of size 50 with internal density 1 produce the same triangle ratio as communities of size 5000 with internal density 1/10. Figure 3 illustrates the range of \( s \) and \( q \) that achieve various triangle and four-cycle ratios. Note that it is possible to achieve \( R_3 = 3 \) and \( R_4 \in \{100, 50, 25\} \) but not \( R_3 = 3 \) and \( R_4 \in \{3, 10\} \).

Next, we show that for almost all achievable pairs of triangle and four-cycle ratios, there exists a ROC construction that matches both ratios asymptotically.

Theorem 3. The ROC model approximates most pairs of triangle and four-cycle ratios.

1. If there exists a graph \( H \) with \( R_3(H) = r_3 \) and \( R_4(H) = r_4 \), then \( 3r_3(3r_3 - 1) \leq 4r_4 \).

2. For any \( r_3 \) and \( r_4 \) such that \( 9r_3^2 \leq 4r_4 \), and \( d = o(n^{1/3}) \), the random graph

\[
G \sim ROC \left(n, d, \frac{16r_3^2}{27r_3^3}, \frac{9r_3^2}{4r_4}\right)
\]

has

\[
\lim_{n \to \infty} \overline{R}_3(G) = r_3 \quad \text{and} \quad \lim_{n \to \infty} \overline{R}_4(G) = r_4.
\]

For every graph with triangle and four-cycle ratios in the narrow range \( 3r_3(3r_3 - 1) \leq 4r_4 \leq 9r_3^2 \), there exists a ROC construction that matches \( r_3 \) and can approximate \( r_4 \) by \( 9r_3^2 \), i.e., up to an additive error \( 3r_3/4 \) (or multiplicative error of at most \( 1/(3r_3 - 1) \) which goes to zero as \( r_3 \) increases).
3.2 Clustering coefficient.

Theorem 4 gives an approximation of the expected clustering coefficient when the degree and average number of communities per vertex grow with \( n \). The exact statement is given in Lemma 17 of Section C, and bounds in a more general setting are given by (4).

**Theorem 4.** Let \( C(v) \) denote the clustering coefficient of a vertex \( v \) with degree at least 2 in a graph drawn from \( \text{ROC}(n,d,s,q) \) with \( d = o(\sqrt{n}) \), \( d < (s-1)qe^sq \), \( d = \omega(sq \log \frac{n}{s}) \), \( s^2q = \omega(1) \), and \( sq = o(d) \). Then

\[
E[C(v)] = (1 + o(1)) \frac{sq^2}{d}.
\]

Unlike in E-R graphs in which local clustering coefficient is independent of degree, higher degree vertices in ROC graphs have lower clustering coefficient. High degree vertices tend to be in more communities, and thus the probability two randomly selected neighbors are in the same community is lower. Figure 4 illustrates the relationship between degree and clustering coefficient, the degree distribution, and the clustering coefficient for two ROC graphs with different parameters and the E-R random graph of the same density.

**Theorem 5.** Let \( C(v) \) denote the clustering coefficient of a vertex \( v \) in a graph drawn from \( \text{ROC}(n,d,s,q) \) with \( d = o(\sqrt{n}) \), \( s = \omega(1) \) and \( \text{deg}(v) \geq 2sq \). Then

\[
E[C(v) \mid \text{deg}(v) = r] = \frac{sq^2}{r} (1 + o_r(1))
\]

**Remark 6.** The dependence between degree and clustering coefficient is the result of the variation in the numbers of communities a vertex is part of. To eliminate this variation and obtain a clustering coefficient distribution that is not highly dependent on degree, we can modify the ROC construction as follows. Instead of selecting \( s \) vertices uniformly at random to make up a community in each step, pre-assign each vertex to precisely \( \frac{d}{sq} \) communities of size \( s \). In this setting the expected clustering coefficient can easily be computed:

\[
E[C(v)] = \Pr[\text{two randomly selected nhbs are from the same community}] = \frac{sq^2}{d}.
\]

Note also, that this variant of the ROC model will produce graphs with fewer isolated vertices.
The mean clustering coefficients are 0.00270, 0.06266, and 0.01595 respectively.

4 Diverse degree distributions and the DROC model

In this section we introduce an extension of our model which produces graphs that match a target degree distribution in expectation. The extension is inspired by the Chung-Lu configuration model: given a degree sequence \(d_1, \ldots, d_n\), an edge is added between each pair of vertices \(v_i\) and \(v_j\) with probability \(\frac{d_id_j}{\sum_{i=1}^{n} d_i}\), yielding a graph where the expected degree of vertex \(v_i\) is \(d_i\) \cite{LuCh}. In the DROC model, a modified Chung-Lu random graph is placed instead of an E-R random graph in each iteration. Instead of normalizing the probability an edge is selected in a community by the sum of the degrees in the community, the normalization constant is the expected sum of the degrees in the community. We use \(D\) to denote a target degree sequence \(t(v_1), \ldots, t(v_n)\), and \(d\) to denote the mean.

\[
\text{DROC}(n, D, s, q).
\]

Output: a graph on \(n\) vertices where vertex \(v_i\) has expected degree \(t(v_i)\).

Repeat \(n/((s-1)q)\) times:

1. Pick a random subset \(S\) of vertices (from \(\{1, 2, \ldots, n\}\)) by selecting each vertex with probability \(s/n\).

2. Add a modified C-L random graph on \(S\), i.e., for each pair in \(S\), add the edge between them independently with probability \(\frac{g(t(v_i))t(v_j)}{sd}\); if the edge already exists, do nothing.

**Theorem 7.** Given a degree distribution \(D\) with mean \(d\) and \(\max_i t(v_i)^2 \leq \frac{sd}{q}\), \(\text{DROC}(n, D, s, q)\) yields a graph where vertex \(v_i\) has expected degree \(t(v_i)\).

We require \(\max_i t(v_i)^2 \leq \frac{sd}{q}\) to ensure that the probability each edge is chosen is at most 1.
Remark 8. Instead of requiring a sequence of \( n \) target degrees as input to the DROC model, we can define the model with a distribution \( D \) of target degrees. In this altered version, Step 0 of the algorithm is to select a target degree for each vertex according to \( D \).

Remark 9. Taking the distribution \( D \) with \( t(v) = d \) for all \( v \) in the DROC model does not yield \( ROC(n, d, s, q) \). The model \( DROC(n, D, s, q) \) is equivalent to \( ROC(n, d, s, \frac{sd}{q}) \).

The following corollary shows that it is possible to achieve a power law degree distribution with the DROC model for power law parameter \( \gamma > 2 \). We use \( \zeta(\gamma) = \sum_{n=1}^{\infty} n^{-\gamma} \) to denote the Riemann zeta function.

Corollary 10. Let \( D \sim D_\gamma \) be the power law degree distribution defined as follows:

\[
Pr[t(v_i) = k] = \frac{k^{-\gamma}}{\zeta(\gamma)},
\]

for all \( 1 \leq i \leq n \). If \( \gamma > 2 \) and

\[
\frac{s}{q} = \omega(1) \frac{\zeta(\gamma)}{\zeta(\gamma - 1)} n^{\frac{1}{\gamma - 1}},
\]

then with high probability \( D \) satisfies the conditions of Theorem 7 and therefore can be used to produce a DROC graph.

4.1 Clustering Coefficient.

We show that by varying \( s \) and \( q \) we can control the clustering coefficient of a DROC graph.

Theorem 11. Let \( C(v) \) denote the clustering coefficient of a vertex \( v \) in graph drawn from \( DROC(n, D, s, q) \) with \( \max t(v_i)^2 \leq \frac{sd}{q} \), \( s = \omega(1) \), \( s/n = o(q) \), and \( t = t(v) \). Then

\[
E[C(v)] = (1 + o(1)) \left( \sum_{u \in V} t(u)^2 \right)^2 \left( (1 - e^{-t})^2 q^2 + c t q^3 \right),
\]

where \( c_t \in [0, 6.2) \) is a constant depending on \( t \).

Equation (10) in the proof of the theorem gives a precise statement of the expected clustering coefficient conditioned on community membership.

5 Discussion and open questions

Modeling real-world graphs. The ROC model captures the degree distribution and clustering coefficient of graphs simultaneously. Previous work [12], [22], and [24] provides models that produce power law graphs with high clustering coefficients. Their results are limited in that the resulting graphs are restricted to a limited range of power-law parameters, and are either deterministic or only analyzable empirically. In contrast, the DROC model is a fully random model designed for a variety of degree distributions (including power law with parameter \( \gamma > 2 \)) and can provably produce graphs with a wide range of clustering coefficient.

Our model therefore may be a useful tool for approximating large graphs. It is often not possible to test algorithms on graphs with billions of vertices (such as the brain, social graphs,
and the internet). Instead, one could use the DROC model to generate a smaller graph with same clustering coefficient and degree distribution as the large graph, and then optimize the algorithm in this testable setting. Further study of such a small graph approximation could provide insight into the structure of the large graph of interest.

Modeling a graph as the union of relatively dense communities has explanatory value for many real-world settings, in particular for social and biological networks. Social networks can naturally be thought of as the union of communities where each community represents a shared interest or experience (i.e. school, work, or a particular hobby); the conceptualization of social networks as overlapping communities has been studied in [23], [29]. Protein-protein interaction networks can also be modeled by overlapping communities, each representing a group of proteins that interact with each other in order to perform a specific cellular process. Analyses of such networks show proteins are involved in multiple cellular processes, and therefore overlapping communities define the structure of the underlying graph [1], [14], [6].

ROC vs mixed membership stochastic block models. Mixed membership stochastic block models have traditionally been applied in settings with overlapping communities [3], [13], [4]. The ROC model differs in two key ways. First, unlike low-rank mixed membership stochastic block models, the ROC model can produce sparse graphs with high triangle and four-cycle ratios. As discussed in the introduction, the over-representation of particular motifs in a graph is thought to be fundamental for its function, and therefore modeling this aspect of local structure is important. Second, in a stochastic block model the size and density of each community and the density between communities are all specified by the model. As a result, the size of the stochastic block model must grow with the number of communities, but the ROC model maintains a succinct description. This observation suggests the ROC model may be better suited for graphs in which there are many communities that are similar in structure, whereas the stochastic block model is better suited for graphs with a small number of communities with fundamentally different structures. Below we discuss extensions of the ROC model that maintain a succinct description and produce more diverse community structures.

Open questions.

1. Consider the following extension. Instead of adding communities of size $s$ and density $q$, we define a probability distribution on a set of pairs $(s_i, q_i)$, and in each iteration choose a pair of parameters $(s_i, q_i)$ from the distribution and build the community $G_{s_i, q_i}$ on $s_i$ randomly selected vertices. Does this modification provide a better approximation for real-world graphs?

2. A further generalization involves adding particular subgraphs from a specified set according to some distribution instead of E-R graphs in each step (e.g., perfect matchings or Hamiltonian paths). Does doing so allow for greater flexibility in tuning the number of various types of motifs present (not just triangles and four-cycles)?

3. A fundamental question in the study of graphs is how to identify relatively dense clusters. For example, clustering protein-protein interaction networks is a useful technique for identifying possible cellular functions of proteins whose functions were otherwise unknown [26], [14]. An algorithm designed specifically to identify the communities in a graph drawn from the ROC
model has potential to become a state-of-the-art algorithm for clustering real-world networks with overlapping community structure.

4. The asymptotic thresholds for properties of E-R graphs have been studied extensively, see [11] for a survey. Such questions are yet to be explored on ROC graphs, e.g., does every nontrivial monotone property have a sharp threshold?

5. How do graph algorithms behave on ROC graphs? For instance, what is the covertime of a random walk on a ROC graph?

References

[1] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. Nature, 466(7307):761–764, 2010.

[2] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services, 2007.

[3] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. Journal of Machine Learning Research, 9(Sep):1981–2014, 2008.

[4] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, Eric P Xing, and Tommi Jaakkola. Mixed membership stochastic block models for relational data with application to protein-protein interactions. In Proceedings of the international biometrics society annual meeting, pages 1–34, 2006.

[5] Uri Alon. Network motifs: theory and experimental approaches. Nature Reviews Genetics, 8(6):450–461, 2007.

[6] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics, 4(1):2, 2003.

[7] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. Science, 286(5439):509–512, 1999.

[8] Fan Chung and Linyuan Lu. Connected components in random graphs with given expected degree sequences. Annals of Combinatorics, 6(2):125–145, 2002.

[9] Paul Erdős and Alfréd Rényi. On random graphs i. Publ. Math. Debrecen, 6:290–297, 1959.

[10] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci, 5(1):17–60, 1960.

[11] Alan Frieze and MichalKaroński. Introduction to Random Graphs. Cambridge University Press, 2015.

[12] Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. Physical review E, 65(2), 2002.
[13] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. Physical Review E, 83(1):016107, 2011.

[14] Nevan J Krogan, Gerard Cagney, Haiyuan Yu, Gouqing Zhong, Xinghua Guo, Alexandr Ignatchenko, Joyce Li, Shuye Pu, Nira Datta, Aaron P Tikuisis, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature, 440(7084):637, 2006.

[15] László Lovász. Large Networks and Graph Limits., volume 60 of Colloquium Publications. American Mathematical Society, 2012.

[16] Priya Mahadevan, Dmitri Krioukov, Marina Fomenkov, Xenofontas Dimitropoulos, Amin Vahdat, et al. The internet as-level topology: three data sources and one definitive metric. ACM SIGCOMM Computer Communication Review, 36(1):17–26, 2006.

[17] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. Science, 298(5594):824–827, 2002.

[18] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks, 2007.

[19] Mark EJ Newman. The structure and function of complex networks. SIAM review, 45(2):167–256, 2003.

[20] Mark EJ Newman. Random graphs as models of networks, pages 35–68. Wiley-VCH Verlag GmbH Co. KGaA, 2005.

[21] Mark EJ Newman. Random graphs with clustering. Physical review letters, 103(5):058701, 2009.

[22] Liudmila Ostroumova, Alexander Ryabchenko, and Egor Samosvat. Generalized preferential attachment: Tunable power-law degree distribution and clustering coefficient. In WAW, pages 185–202. Springer, 2013.

[23] Gergely Palla, Albert-László Barabási, and Tamás Vicsek. Quantifying social group evolution. arXiv preprint arXiv:0704.0744, 2007.

[24] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. Science, 297(5586):1551–1555, 2002.

[25] Sen Song, Per Jesper Sjöström, Markus Reigl, Sacha Nelson, and Dmitri B Chklovskii. Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS biology, 3(3):e68, 2005.

[26] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, et al. A human protein-protein interaction network: a resource for annotating the proteome. Cell, 122(6):957–968, 2005.
[27] Steven H. Strogatz. Exploring complex networks. Nature, 410(6825):268–276, 2001.

[28] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of “small-world” networks. Nature, 393(6684):440–442, 1998.

[29] Jierui Xie, Boleslaw K Szymanski, and Xiaoming Liu. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, pages 344–349. IEEE, 2011.

[30] Esti Yeger-Lotem, Shmuel Sattath, Nadav Kashtan, Shalev Itzkovitz, Ron Milo, Ron Y Pinter, Uri Alon, and Hanah Margalit. Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. Proceedings of the National Academy of Sciences of the United States of America, 101(16):5934–5939, 2004.
A  Limitations of previous approaches

**Theorem 12.** Let $G$ be a graph on $n$ vertices obtained by repeatedly adding triangles on sets of three randomly chosen vertices. If the average degree is less than $\sqrt{n}$, the expected ratio of triangles to edges is at most $2/3$.

**Proof.** Let $t$ be the number of triangles added and $d$ the average degree, so $d = 6t/n$. To ensure that $d < \sqrt{n}$, $t < n^{3/2}/6$. The total number of triangles in the graph is $t + (d/n)^3 \binom{n}{3} = t + d^3/6 = t + 36t^3/n^3$. It follows that the expected ratio of triangles to edges is at most

$$\frac{t + 36 \left(\frac{t}{n}\right)^3}{3t} \leq \frac{2}{3}.$$ 

\[\square\]

**Proof.** (of Theorem 1) Let $\sigma_1 \ldots \sigma_{\text{rank}(M)}$ denote the eigenvalues of $M$.

$$E[\#k\text{-cycles}] = \sum_{i_1 \neq i_2 \ldots \neq i_k} M_{i_1i_2}M_{i_2i_3} \ldots M_{i_ki_1} \leq Tr(M^k) = \sum_{i=1}^{\text{rank}(M)} \sigma_i^k \leq \text{rank}(M)d^k.$$ 

\[\square\]

B  Connectivity of the ROC model

We describe the thresholds for connectivity for $ROC(n,d,s,q)$ networks. A vertex is isolated if it is has no adjacent edges. A community is isolated if it does not intersect any other communities.

Here we use the abbreviation a.a.s. for asymptotically almost surely. An event $A_n$ happens a.a.s. if $\Pr[A_n] \to 1$ as $n \to \infty$.

**Theorem 13.** For $(s - 1)q(\ln n + c) \leq d \leq (s - 1)qe^{\delta q}(1 - \varepsilon)$, a graph from $ROC(n,d,s,q)$ a.a.s. has at most $(1 + o(1)) \frac{e^{-\varepsilon}}{1-\varepsilon}$ isolated vertices.
Proof. We begin by computing the probability a vertex is isolated,

\[
\Pr[v \text{ is isolated}] = \sum_{i=0}^{\frac{nd}{s^2q}} \Pr[v \text{ is in } i \text{ communities}][(1-q)^i]
\]

\[
= (1 + o(1)) \sum_{i=1}^{\frac{nd}{s^2q}} \left( \frac{\frac{nd}{s(s-1)q}}{i} \right) \left( \frac{s}{n} \right)^i \left( 1 - \frac{s}{n} \right)^{\frac{nd}{s(s-1)q} - i} e^{-sqi}
\]

\[
\leq (1 + o(1)) e^{-\frac{d}{s(s-1)q}} \sum_{i=0}^{\frac{nd}{s^2q}} \left( \frac{de^{-sq} + \frac{s}{n}}{(s-1)q} \right)^i
\]

\[
= (1 + o(1)) e^{-\frac{d}{s(s-1)q}} \sum_{i=1}^{\frac{nd}{s^2q}} \left( \frac{de^{-sq}}{(s-1)q} \right)^i
\]

\[
= (1 + o(1)) \left( e^{-\frac{d}{s(s-1)q}} \right) \left( \frac{1}{1 - \varepsilon} \right).
\]

Let \( X \) be a random variable that represents the number of isolated vertices of a graph drawn from \( ROC(n,d,s,q) \). We compute

\[
\Pr[X > 0] \leq \mathbb{E}[X] = (1 + o(1)) n \left( e^{-\frac{d}{s(s-1)q}} \right) \left( \frac{1}{1 - \varepsilon} \right) = (1 + o(1)) \left( \frac{e^{-c}}{1 - \varepsilon} \right).
\]

\[\square\]

**Theorem 14.** A graph from \( ROC(n,d,s,q) \) with \( s = o(\sqrt{n}) \) has no isolated communities a.a.s. if

\[
\frac{d}{q} > \log \frac{nd}{s^2q}.
\]

**Proof.** We construct a “community graph” and apply the classic result that \( G(n,p) \) will a.a.s. have no isolated vertices when \( p > (1+\varepsilon) \log n/n \) for any \( \varepsilon > 0 \)[9]. In the “community graph” each vertex is a community and there is an edge between two communities if they share at least one vertex; a ROC graph has no isolated communities if and only if the corresponding “community graph” is connected. The probability two communities don’t share a vertex is \( 1 - \varepsilon \). Since communities are selected independently, the “community graph” is an instance of \( G \left( \frac{\frac{nd}{s(s-1)q}}{1 - (\frac{s}{n})^s}, 1 - e^{s^2/n} \right) \). By the classic result, approximating the parameters by \( \frac{\frac{nd}{s^2q}}{1 - e^{s^2/n}}, 1 - e^{s^2/n} \), this graph is connected when

\[
1 - e^{-s^2/n} > \frac{\log \frac{nd}{s^2q}}{\frac{nd}{s^2q}}.
\]

Since \( s = o(\sqrt{n}) \) is small, the left side of the inequality is approximately \( s^2/n \), yielding the equivalent statement

\[
\frac{d}{q} > \log \frac{nd}{s^2q}.
\]

\[\square\]
Note that the threshold for isolated vertices is higher, meaning that if a ROC graph a.a.s has no isolated vertices, then it a.a.s has no isolated communities. These two properties together imply the graph is connected.

C Section 2 proofs

Proof. (of Lemma 2) Let $G \sim ROC(n,d,s,q)$ and $u,v \in V(G)$. First note that without information about whether $u$ and $v$ are in community together $\Pr[u \sim v] = d/n = o(1)$ because each edge is equally likely. However, $\Pr[u \sim v \mid u,v$ are in a common community $] = q+o(1)$. We show that both the triangle count and the four-cycle count are dominated by cycles contained entirely in one community.

We compute $\mathbb{E}[C_3(G)]$ by counting the total number of triangles in $G$. Let $T_1$ be the number triangles with all three edges originating in one community, $T_2$ be the number of triangles with two edges originating in the same community and the third edge originating in a different community, and $T_3$ be the number of triangles with edges originating in three different communities. We compute

$$\mathbb{E}[T_1] = (\text{# com.}) \mathbb{E}[\text{triangles in a com.}] = \frac{nd}{s(s-1)q} \frac{s^3q^3}{6} = \frac{ndsq^2}{6} (1 + o(1))$$

$$\mathbb{E}[T_2] = (\text{# com.}) \mathbb{E}[\text{#two paths } u \sim v, v \sim w \text{ in a com.}] \Pr[ u \sim w \text{ in other com.}]$$

$$= \frac{nd}{s(s-1)q} s \frac{s-1}{2} q^2 \frac{d}{n} = \frac{d^2sq^2}{2} (1 + o(1))$$

$$\mathbb{E}[T_3] = \text{(# triples } u,v,w \in V(G) \Pr[u,v,w \text{ form a triangle}] = \left(\frac{n}{3}\right) \left(\frac{d}{n}\right)^3 = \frac{d^3}{6}.$$

Therefore

$$\mathbb{E}[R_3(G)] = \frac{2\mathbb{E}[T_1 + T_2 + T_3]}{nd} = \frac{sq^2}{3} (1 + o(1)).$$

Similarly, we compute $\mathbb{E}[C_4(G)]$ by summing over different categories of four-cycles based on the shared community membership of the vertices. For simplicity suppose the $a,b,c,d$ are the vertices of the four-cycle and let $C_1, \ldots C_4$ denote different communities. If $\{a,b,c,d\} \in C_1$, the cycle is type 1. If $\{a,b,c\} \in C_1$ and $\{a,c,d\} \in C_2$, the the cycle is type 2. If $\{a,b,d\} \in C_1$, $\{b,c\} \in C_2$, $\{c,d\} \in C_3$, then the cycle is type 3. If $\{a,b\} \in C_1$, $\{b,c\} \in C_2$, $\{c,d\} \in C_3$, $\{d,a\} \in C_4$, then the cycle is type
4. Let \( F_i \) be the number of cycles of type \( i \). We compute

\[
\mathbb{E}[F_1] = \binom{n}{2} \left( \frac{2}{n} \right)^4 \left( \frac{nd}{s(s-1)q} \right)^2 \left( \frac{(s-2)q^2}{2} \right)^2 = \frac{8}{n} (1 + o(1))
\]

\[
\mathbb{E}[F_2] = \mathbb{E}[\# vertex pairs \ u, v \text{ in two of the same coms. } | \ (E[#\text{common nhbs of } u, v \text{ in a com.}])^2
\]

\[
\mathbb{E}[F_3] = \binom{n}{3} \left( \frac{2}{n} \right)^4 \left( \frac{nd}{s(s-1)q} \right)^3 \left( \frac{q^2}{2} \right) \left( \frac{d}{n} \right)^2 = \frac{s^3}{2}
\]

\[
\mathbb{E}[F_4] = \binom{n}{4} \left( \frac{d}{n} \right)^4 = \frac{d^4}{8}
\]

Therefore

\[
\mathbb{E}[R_3(G)] = \frac{2\mathbb{E}[F_1 + F_2 + F_3 + F_4]}{nd} = \frac{8}{3} (1 + o(1)).
\]

**Proof.** (of Theorem 3) (1) For each edge in \( H \), let \( t_e \) be the number of triangles containing \( e \), so \( \sum_{e \in E(H)} t_e = 3C_3(H) = 3r_3|E(H)| \). If triangles \( abc \) and \( abd \) are present, then so is the four-cycle \( acbd \). This four-cycle may also be counted via triangles \( cad \) and \( cdb \). Therefore \( C_4(H) \geq \frac{1}{2} \sum_{e \in E(H)} \left( t_e \right) \). This expression is minimized when all \( t_e \) are equal. We therefore obtain

\[
r_4|E(H)| = C_4(H) \geq \frac{|E(H)|}{2} \left( \frac{3r_3}{2} \right) = \frac{3r_3(3r_3 - 1)|E(H)|}{4}.
\]

It follows that \( \frac{3r_3(3r_3 - 1)}{4r^4} \leq 1 \).

(2) Since the hypothesis guarantees \( q \leq 1 \), applying Lemma 2 to \( G \sim ROC \left( n, d, \frac{16r^2}{27r^2_n}, \frac{9r^2}{4r^2} \right) \) implies the desired statements.

**Remark 15.** Theorem 4 gives bounds expected clustering coefficient up to factors of \( (1 + o(1)) \). The clustering coefficient at a vertex is only well-defined if the vertex has degree at least two. Given the assumption in Theorem 4 that \( d = \omega(sq \log \frac{nd}{s}) \), \( d < (s-1)q\epsilon^q \), and \( s = \omega(1) \), Lemma 16 implies that the fraction of vertices of degree strictly less than two is \( o(1) \). Therefore we ignore the contribution of these terms throughout the computations for Theorem 4 and supporting Lemma 17. In addition we divide by \( \text{deg}(v)^2 \) rather than by \( \text{deg}(v)(\text{deg}(v) - 1) \) in the computation of the clustering coefficient since this modification only affects the computations up to a factor of \( (1 + o(1)) \).

**Lemma 16.** If \( d = \omega(sq \log \frac{nd}{s}) \), \( s = \omega(1) \), \( s = o(n) \), and \( d < (s-1)q\epsilon^q \), then a graph from \( ROC(n, d, s, q) \) a.a.s. has no vertices of degree less than 2.
Proof. Theorem [13] implies there are no isolated vertices a.a.s. We begin by computing the probability a vertex has degree one.

\[
\Pr[\deg(v) = 1] = \sum_{i=1}^{\frac{nd}{s^2 q}} \Pr[v \text{ is in } i \text{ communities}] q(1 - q)^{si-1}
\]

\[
= \sum_{i=1}^{\frac{nd}{s^2 q}} \left( \frac{nd}{s(s-1)q} \right)^i \left( \frac{s}{n} \right)^i \left( 1 - \frac{s}{n} \right)^{\frac{nd}{s(s-1)q}-i} q(1 - q)^{si-1}
\]

\[
\leq (1 + o(1)) \sum_{i=1}^{\frac{nd}{s^2 q}} \left( \frac{nd}{s(s-1)q} \right)^i \left( \frac{s}{n} \right)^i e^{-\frac{d}{sq} + \frac{s}{n} qe^{-qsi+q}}
\]

\[
= (1 + o(1)) qe^{-\frac{d}{sq}} \sum_{i=1}^{\frac{nd}{s^2 q}} \left( \frac{de^{-sq}}{(s-1)q} \right)^i
\]

\[
= O \left( \frac{de^{-sq}-\frac{d}{sq}}{s} \right)
\]

Let \( X \) be a random variable that represents the number of degree one vertices of a graph drawn from \( ROC(n,d,s,q) \). When \( d = \omega(sq \log \frac{nd}{s}) \), we obtain

\[
\Pr[X > 0] \leq E[X] = O \left( \frac{nde^{-sq}-\frac{d}{sq}}{s} \right) = o(1).
\]

Lemma 17. Let \( C(v) \) denote the clustering coefficient of a vertex \( v \) of degree at least 2 in a graph drawn from \( ROC(n,d,s,q) \) with \( d = o(\sqrt{n}) \) and \( d = \omega(sq \log \frac{nd}{s}) \). Then

\[
E[C(v)] = (1 + o(1)) \left( \sum_{i=1}^{\frac{nd}{s^2 q}} \left( \frac{nd}{s^2 q} \right)^i \left( \frac{s}{n} \right)^i \left( 1 - \frac{s}{n} \right)^{\frac{nd}{s^2 q}-i} \frac{s(s-1)q^3 k}{(sqk + 2 - 2q)^2} \right).
\]

Proof. For ease of notation, we ignore factors of \((1 + o(1))\) throughout as described in Remark [15].

First we compute the expected clustering coefficient of a vertex from an \( ROC(n,d,s,q) \) graph given \( v \) is contained in precisely \( k \) communities. Let \( X_1, \ldots, X_k \) be random variables representing the degree of \( v \) in each of the communities, \( X_i \sim Bin(s,q) \). We have

\[
E[C(v) | v \text{ in } k \text{ communities}] = E \left[ \frac{\sum_{i=1}^{k} X_i(X_i-1)q}{\left( \sum_{i=1}^{k} X_i \right)^2} \right]
\]

\[
= qk E \left[ \frac{X_1(X_1-1)}{(sq(k-1)+X_1)^2} \right]
\]

\[
= qk E \left[ \frac{X_1^2}{(sq(k-1)+X_1)^2} \right] - qk E \left[ \frac{X_1}{(sq(k-1)+X_1)^2} \right].
\]
Write $X_1 = \sum_{i=1}^s y_i$ where $y_i \sim Bernoulli(q)$. Using linearity of expectation and the independence of the $y_i$s we have

$$E \left[ \frac{X_1}{(sq(k-1) + X_1)^2} \right] = s E \left[ \frac{y_1}{(sq(k-1) + (s-1)q + y_1)^2} \right] = \frac{sq}{(sq(k-1) + (s-1)q + 1)^2},$$

and

$$E \left[ \frac{X_1^2}{(sq(k-1) + X_1)^2} \right] = E \left[ \frac{(\sum_{i=1}^s y_i)^2}{(sq(k-1) + \sum_{i=1}^s y_i)^2} \right]
= s E \left[ \frac{y_1^2}{(sq(k-1) + q(s-1) + y_1)^2} \right] + s(s-1) E \left[ \frac{(y_1y_2)^2}{(sq(k-1) + (s-2)q + y_1 + y_2)^2} \right]
= \frac{sq}{(sq(k-1) + q(s-1) + 1)^2} + \frac{s(s-1)q^2}{(sq(k-1) + (s-2)q + 2)^2}.$$

Substituting in these values into (1), we obtain

$$E[C(v)|v \in k \text{ communities}] = qk \left( \frac{s(s-1)q^2}{(sq(k-1) + (s-2)q + 2)^2} \right) = \frac{s(s-1)q^3k}{(sqk + 2 - 2q)^2}. \tag{2}$$

Let $M$ be the number of communities a vertex is in, so $M \sim Bin \left( \frac{nd}{sq}, \frac{s}{n} \right)$. It follows

$$E[C(v)] = \sum_{i=1}^{\frac{nd}{sq}} \Pr \left[ v \text{ in } k \text{ communities } \right] E[C(v)|v \text{ in } k \text{ communities}]
= \sum_{i=1}^{\frac{nd}{sq}} \binom{s}{i} \left( \frac{s}{n} \right)^i \left( 1 - \frac{s}{n} \right)^{\frac{nd}{sq} - i} \frac{s(s-1)q^3k}{(sqk + 2 - 2q)^2}.$$

The proof of Theorem 4 relies on the follow two lemmas regarding expectation of binomial random variables.

**Lemma 18.** Let $X \sim Bin(n, p)$. Then

1. $E \left[ \frac{1}{X+1} \mid X \geq 1 \right] = \frac{1-(1-p)^{n+1}-(n+1)p(1-p)^n}{p(n+1)}$ and
2. $E \left[ \frac{1}{X+1} \right] = \frac{1-(1-p)^{n+1}}{p(n+1)}$. 

18
Proof. Observe
\[
E \left[ \frac{1}{X+1} \mid X \geq 1 \right] = \sum_{i=1}^{n} \binom{n}{i} \frac{p^i (1-p)^{n-i}}{i+1} = \frac{1}{p(n+1)} \sum_{i=1}^{n} \binom{n+1}{i+1} p^{i+1} (1-p)^{n-i} = \frac{1 - (1-p)^{n+1} - (n+1)p(1-p)^n}{p(n+1)}.
\]

Similarly
\[
E \left[ \frac{1}{X+1} \right] = \sum_{i=0}^{n} \frac{n}{i} \binom{n}{i} \frac{p^i (1-p)^{n-i}}{i+1} = \frac{1}{p(n+1)} \sum_{i=0}^{n} \binom{n+1}{i+1} p^{i+1} (1-p)^{n-i} = \frac{1 - (1-p)^{n+1}}{p(n+1)}.
\]

\[\square\]

**Lemma 19.** Let \( X \sim \text{Bin}(n,p) \). Then
\[
E \left[ \frac{1}{X} \mid X \geq 1 \right] \leq \frac{1}{p(n+1)} \left(1 + \frac{3}{p(n+2)}\right).
\]

Proof. Note that when \( X \geq 1 \),
\[
\frac{1}{X} \leq \frac{1}{X+1} + \frac{3}{(X+1)(X+2)}.
\]

By Lemma 18
\[
E \left[ \frac{1}{X+1} \mid X \geq 1 \right] \leq \frac{1}{p(n+1)}.
\]

We compute
\[
E \left[ \frac{1}{(X+1)(X+2)} \mid X \geq 1 \right] = \sum_{i=1}^{n} \binom{n}{i} \frac{p^i (1-p)^{n-i}}{(i+1)(i+2)} = \frac{1}{p^2(n+2)(n+1)} \sum_{i=1}^{n} \binom{n+2}{i+2} p^{i+2} (1-p)^{n-i} \leq \frac{1}{p^2(n+2)(n+1)}.
\]

Taking expectation of (3) gives
\[
E \left[ \frac{1}{X} \mid X \geq 1 \right] \leq \frac{1}{p(n+1)} \left(1 + \frac{3}{p(n+2)}\right).
\]
\[\square\]
Proof. (of Theorem 4) For ease of notation, we ignore factors of \((1 + o(1))\), as described in Remark 15. It follows from (2) in the proof of Lemma 17 that

\[
\frac{q}{k+1} \leq \mathbb{E}[C(v) | v \in k \text{ communities}] \leq \frac{q}{k},
\]

where the left inequality holds when \(q(s-1) \geq 5\).

We now compute upper and lower bounds on \(\mathbb{E}[C(v)]\), assuming \(v\) is in some community. Let \(M\) be the random variable indicating the number of communities containing \(v\), \(M \sim \text{Bin} \left( \frac{ns}{(s-1)q}, \frac{s}{n} \right)\). It follows

\[
\mathbb{E}[C(v)] = \sum_{k=1}^{\frac{nd}{s(s-1)q}} \mathbb{P}[M = k] \mathbb{E}[C(v) | M = k]
\]

\[
q \mathbb{E} \left[ \frac{1}{M+1} \mid M \geq 1 \right] \leq \mathbb{E}[C(v)] \leq q \mathbb{E} \left[ \frac{1}{M} \mid M \geq 1 \right].
\]

Applying Lemmas 18 and 19 to the lower and upper bounds respectively, we obtain

\[
q \left( 1 - \left(1 - \frac{s}{n} \right) \frac{nd}{s(s-1)q} + 1 \right) \left(1 - \frac{s}{n} \right) \frac{nd}{s(s-1)q} \right) \leq \mathbb{E}[C(v)] \leq q \left( \frac{d}{(s-1)q} + \frac{s}{n} \right) \left(1 + \frac{3}{1 + \frac{d}{(s-1)q} + \frac{2s}{n}}\right)
\]

which for \(s = o(n)\) simplifies to

\[
(1 + o(1)) \frac{(s-1)q^2}{d} \left(1 - \frac{nd}{s(s-1)q} e^{-d/(s-1)q} \right) \leq \mathbb{E}[C(v)] \leq \frac{(s-1)q^2}{d} \left(1 + \frac{(s-1)q}{d} \right) (1 + o(1)) .
\]

Under the assumptions \(s^2q = \omega(1)\) and \(sq = o(d)\), we obtain our desired result

\[
\mathbb{E}[C(v)] = (1 + o(1)) \left(\frac{sq^2}{d}\right).
\]

The following lemma will be used in the proof of Theorem 5.

**Lemma 20.** The \(X\) be a nonnegative integer drawn from the discrete distribution with density proportional to \(f(x) = x^{r-x}e^{-ax}\). Let \(z = \text{arg max } f(x)\). Then

\[
\Pr[|x - z| \geq 2t\sqrt{z}] \leq e^{-t^2}.
\]

**Proof.** First we observe that \(f\) is logconcave:

\[
\frac{d^2}{dx^2} \ln f(x) = \frac{d}{dx} \left(\frac{d}{x} + \frac{r}{x} - 1 - \ln x\right) = -\frac{r}{x^2} - \frac{1}{x}
\]
which is nonpositive for all \( x \geq 0 \). We will next bound the standard deviation of this density, so that we can use an exponential tail bound for logconcave densities. To this end, we estimate \( \max f \). Setting its derivative to zero, we see that at the maximum, we have

\[
a + 1 = \frac{r}{x} - \ln x.
\]  

(5)

The maximizer \( z \) is very close to

\[
\frac{r}{(a + 1) + \ln(r/(a + 1))},
\]

(6)

and the maximum value \( z \) satisfies \( z^{r-z}e^{-az} = z^r e^{-r+z} \). Now we consider the point \( z + \delta \) where \( f(z + \delta) = f(z)/e \), i.e.,

\[
(z + \delta)^{r-z-\delta}e^{-az-a\delta} \leq e^{-1}.
\]

The LHS is

\[
\left(1 + \frac{\delta}{z}\right)^{r-z} z^{-\delta} \left(1 + \frac{\delta}{z}\right)^{-\delta} e^{-a\delta} \leq e^{\delta(z-1-a-\ln z)}e^{-\frac{\delta^2}{2}}.
\]

where in the second step we used the optimality condition (5). Thus for \( \delta = (1 + o(1)) \sqrt{z} \),

\[
f(x + \delta) = f \left( \left(1 - \frac{1}{t}\right)x + \frac{1}{t}(x + t\delta) \right) \geq f(x)^{1-1/t} f(x + t\delta)^{1/t}
\]

for any \( t \geq 1 \). It follows

\[
f(x + t\delta) \leq f(x)/e^t
\]

(7)

for all \( t \) (since we can apply the same argument for \( z - \delta \)). Taking \( x = z \) in (7) and using the observation \( \sum_{x \in \mathbb{Z}^+} f(x) \geq f(z) \), it follows that

\[
\Pr[x = z + t\sqrt{z}] \leq e^{-t} \quad \text{and} \quad \Pr[x = z - t\sqrt{z}] \leq e^{-t}
\]

and so

\[
\Pr[|x - z| \geq t\sqrt{z}] \leq 2e^{-t} \leq e^{-t+1}.
\]

Proof. (of Theorem 5). Let \( M \) denote the number of communities a vertex \( v \) is selected to participate in. We can write

\[
\mathbb{E}[C(v)|\deg(v) = r] = \sum_{k=r}^{r} \mathbb{E}[C(v)|\deg(v) = r, M = k] \Pr[M = k|\deg(v) = r]
\]

\[
= \sum_{k=1}^{r} \mathbb{E}[C(v)|\deg(v) = r, M = k] \Pr[\deg(v) = r|M = k] \Pr[M = k]\Pr[\deg(v) = r].
\]

1which says that for any \( x, y \) and any \( \lambda \in [0, 1] \), we have \( f(\lambda x + (1 - \lambda)y) \geq f(x)^{\lambda} f(y)^{1-\lambda} \).
First we compute the expected clustering coefficient of a degree $r$ vertex given that it is $k$ communities:

$$E[C(v)|\text{deg}(v) = r \text{ and } M = k] = \frac{\sum_{i \neq j, i, j \in N(v)} q(\text{Pr}[i, j \text{ part of same community}])}{\text{deg}(v)(\text{deg}(v) - 1)} = \frac{q}{k}.$$ 

Next we note that $M$ is a drawn from a binomial distribution, and the degree of $v$ is drawn from a sum of $k$ binomials, each being $\text{Bin}(s,q)$. Therefore,

$$\text{Pr}[M = k] \text{Pr}[	ext{deg}(v) = r|M = k] = \left(\frac{nd}{k}\right) \left(\frac{s}{n}\right)^k \left(1 - \frac{s}{n}\right)^{\frac{nd}{s-1} - k} \left(\frac{sk}{r}\right)^r (1 - q)^{sk - r}.$$ 

Using this we obtain

$$E[C(v)|\text{deg}(v) = r] = \frac{\sum_{k=1/2}^{\infty} k \cdot \text{Pr}[M = k] \text{Pr}[\text{deg}(v) = r|M = k]}{\sum_{k=1/2}^{\infty} \text{Pr}[M = k] \text{Pr}[\text{deg}(v) = r|M = k]}$$

$$= \left(1 + o(1)\right) q \frac{\sum_{k=1/2}^{\infty} k \cdot \left(\frac{d}{(s-1)q}\right)^k e^{-\frac{d}{(s-1)q} + \frac{sk}{r}} \left(\frac{skq}{r}\right)^r e^{-skq + qr}}{\sum_{k=1/2}^{\infty} \left(\frac{d}{(s-1)q}\right)^k k^{r-k} e^{-skq}}.$$

Therefore, (8) is the same as $qE[1/x]$ when $x$ is a nonnegative integer drawn from the discrete distribution with density proportional to $f(x) = x^{r-x}e^{-ax}$. We let $z$ be as in [6] of Lemma 20, so $z \approx \frac{r}{sq}$. We use Lemma 20 to bound

$$E\left[\frac{1}{x} - \frac{1}{z}\right] \leq \sum_{t=1}^{\infty} \left(\frac{1}{z} - \frac{1}{z + t\sqrt{z}}\right) e^{-t} + \sum_{t=1}^{\sqrt{z}-1} \left(\frac{1}{z - t\sqrt{z}} - \frac{1}{z}\right) e^{-t}$$

$$= \sum_{t=1}^{\infty} \frac{t\sqrt{z}e^{-t}}{z(z + t\sqrt{z})} + \sum_{t=1}^{\sqrt{z}-1} \frac{t\sqrt{z}e^{-t}}{z(z - t\sqrt{z})}$$

$$\leq \frac{1}{z} \sum_{t=1}^{\infty} \frac{te^{-t}}{\sqrt{z} + 1} + \frac{\sqrt{z}}{z} \left(\sum_{t=1}^{\sqrt{z}/3} 3te^{-t} + \sum_{t=\sqrt{z}/3}^{\sqrt{z}-1} te^{-t}\right)$$

$$= O(1) + O(1) + O\left(\frac{\sqrt{z}}{3} e^{-\frac{\sqrt{z}}{3}}\right) = O\left(\frac{1}{z\sqrt{z}}\right).$$
Using this and approximating $z$ by $\frac{qs}{r^2}$, the expectation of $x$ with respect to the density proportional to $f$ can be estimated:

\[
q E\left[\frac{1}{x}\right] = \frac{q}{z} \left(1 + O\left(\frac{1}{\sqrt{z}}\right)\right) = (1 + o(1)) \frac{qs^2}{r} \left(1 + O\left(\sqrt{\frac{sq}{r}}\right)\right) = (1 + o_r(1)) \frac{qs^2}{r}
\]

as claimed.

\[\square\]

### D Section 4 proofs

**Proof.** (of Theorem 11) Let $v$ be a vertex with target degree $t = t(v)$, and let $k$ denote the number communities containing $v$. First we claim $\deg(v) \sim Bin((s-1)k, \frac{tq}{s})$. Let $s$ be an arbitrary vertex of a community $S$ containing $v$.

\[
Pr[s \sim v \text{ in } S] = \sum_{u \in V} Pr[s = u] Pr[v \sim u \text{ in } S] = \sum_{u \in V} \frac{1}{n} \frac{t(u)q}{ds} = \frac{tq}{s}.
\]

A vertex in $k$ communities has the potential to be adjacent to $(s-1)k$ other vertices, and each adjacency occurs with probability $tq/s$.

Next, let $N_u$ be the event that a randomly selected neighbor of vertex $v$ is vertex $u$. We compute

\[
Pr[N_u] = \sum_r \frac{Pr[u \sim v \mid \deg(v) = r] Pr[\deg(v) = r]}{r} = \sum_r \frac{Pr[u \sim v \mid \deg(v) = r] Pr[\deg(v) = r]}{r}
\]

\[
= Pr[u \sim v] \frac{1}{\deg(v)} \left| \frac{1}{X+1} \right|
\]

\[
= (1 + o(1)) \left(\frac{s}{n}\right)^2 \frac{n}{(s-1)q} \frac{t(u)q}{sd} \left(1 - e^{-tk}\right)
\]

\[
= (1 + o(1)) \frac{t(u)q}{kdn} (1 - e^{-tqk}).
\]

To see (9), note that by the first claim $E\left[\frac{1}{\deg(v) \mid u \sim v}\right] = E\left[\frac{1}{X+1}\right]$ where $X \sim Bin((s-1)k-1, \frac{tq}{s})$. Applying Lemma 18 and assuming $s = \omega(1)$, we obtain

\[
E\left[\frac{1}{\deg(v) \mid u \sim v}\right] = \frac{1 - \frac{(s-1)k}{(s-1)k + \frac{tq}{s}}}{(s-1)k + \frac{tq}{s}} = (1 + o(1)) \frac{1 - e^{-tqk}}{tkq}.
\]

Now we compute the expected clustering coefficient conditioned on the number of communities.
the vertex is part of under the assumption that \( s/n = o(q) \). Observe

\[
E[C(v) \mid v \text{ in } k \text{ communities}] = \sum_{u,w} N_u N_w \Pr[u \sim w \mid u \sim v \text{ and } w \sim v]
\]

\[
= \sum_{u,w} t(u)t(w) \frac{(1 - e^{-tqk})^2}{(qkd'n)^2} \left( \frac{1}{k} + \frac{s}{n} \right) \frac{n}{(s-1)q} \frac{t(u)t(w)q}{sd}
\]

\[
= (1 + o(1)) \frac{(1 - e^{-tqk})^2 \left( \sum_{u \in V} t(u)^2 \right)^2}{qd^3k^3n^2s}.
\] (10)

Next compute the expected clustering coefficient without conditioning on the number of communities. To do so we need to compute the expected value of the function \( f(k) = \frac{(1 - e^{-kq})^2}{k^3} \). We first use Taylor’s theorem to give bounds on \( f(k) \). For all \( k \), there exists some \( z \in [1/q, k] \) such that

\[
f(k) = f \left( \frac{1}{q} \right) + f' \left( \frac{1}{q} \right) \left( k - \frac{1}{q} \right) + \frac{f''(z)}{2} \left( k - \frac{1}{q} \right)^2.
\]

Note that for \( z \in [1/q, k] \)

\[
f''(z) = \frac{12(1 - e^{-kqt})^2}{k^5} - \frac{12e^{-kqt}(1 - e^{-kqt})q^2t^2}{k^4} - \frac{2e^{-kqt}(1 - e^{-kqt})q^2t^2}{k^3} \leq \frac{12(1 - e^{-kqt})^2}{k^5} + \frac{2e^{-kqt}q^2t^2}{k^3} \leq q^5 \left( 12 + 2t^2e^{-2t} \right),
\]

and

\[
f''(z) \geq 0.
\]

It follows that

\[
f \left( \frac{1}{q} \right) + f' \left( \frac{1}{q} \right) \left( k - \frac{1}{q} \right) \leq f(k) \leq f \left( \frac{1}{q} \right) + f' \left( \frac{1}{q} \right) \left( k - \frac{1}{q} \right) + q^5 \left( 6 + t^2e^{-2t} \right) \left( k - \frac{1}{q} \right)^2. \] (11)

Let \( M \sim Bin(n/(sq), s/n) \) be the random variable for the number of communities a vertex \( v \) is part of. (Since \( s = \omega(1) \) replacing the number of communities by \( n/(sq) \) changes the result by a factor of \( (1 + o(1)) \).) We use [11] to give bounds on the expectation of \( f(M) \),

\[
E[f(M)] \leq E \left[ f \left( \frac{1}{q} \right) + f' \left( \frac{1}{q} \right) \left( M - \frac{1}{q} \right) + q^5 \left( 12 + 2t^2e^{-2t} \right) \left( M - \frac{1}{q} \right)^2 \right]
\]

\[
= (1 - e^{-t})^2q^3 + \frac{1}{q} \left( 1 - \frac{s}{n} \right) q^5 \left( 6 + t^2e^{-2t} \right)
\]

\[
\leq (1 - e^{-t})^2q^3 + q^4 \left( 6 + t^2e^{-2t} \right)
\]

and

\[
E[f(M)] \geq E \left[ f \left( \frac{1}{q} \right) + f' \left( \frac{1}{q} \right) \left( M - \frac{1}{q} \right) \right] = (1 - e^{-t})^2q^3.
\]
Therefore $E[f(M)] = (1 - e^{-t})^2 q^3 + c_t q^4$ for some constant $c_t \in [0, 6.2)$.

Finally, we compute

$$E[C(v)] = \sum_k \Pr[M = k]\frac{(1 - e^{-tk})^2}{qd^3 k^3 n^2 s} \left(\sum_{u \in V} l(u)^2\right)^2$$

$$= \frac{(\sum_{u \in V} l(u)^2)^2}{qd^3 n^2 s} E[f(M)]$$

$$= (1 + o(1)) \frac{(\sum_{u \in V} l(u)^2)^2}{d^3 n^2 s} ((1 - e^{-t})^2 q^2 + c_t q^3).$$

Proof. (of Corollary 10) Let $d = \text{mean}(D)$. We compute

$$E[d] = \sum_{k=1}^{\infty} \frac{k-\gamma+1}{\zeta(\gamma)} = \frac{\zeta(\gamma - 1)}{\zeta(\gamma)}.$$

Next we claim that with high probability the maximum target degree of a vertex is at most $t_0 = n^{2/(\gamma - 1)}$. Let $X$ be the random variable for the number of indices $i$ with $t(v_i) > k_0$.

$$\Pr\left[ \max_{i} t(v_i) > t_0 \right] \leq E[X] = n \Pr[t(v_1) > t_0] \leq n \sum_{i=0+1}^{\infty} \frac{i^{-\gamma}}{\zeta(\gamma)}$$

$$\leq n \int_{i=t_0}^{\infty} \frac{i^{-\gamma}}{\zeta(\gamma)} = \left(\frac{1}{\zeta(\gamma)(\gamma - 1)}\right) n t_0^{1-\gamma} = o(1).$$

It follows that $\max_i t(v_i)^2 \leq n^{\frac{1}{\gamma - 1}}$, and so $\max_i t(v_i)^2 \leq \frac{sd}{q}$. □