Approximating Sparse Graphs: The Random Overlapping Communities Model

Samantha Petti[*] Santosh S. Vempala[†]

February 13, 2018

Abstract

How can we approximate sparse graphs and sequences of sparse graphs (with average degree unbounded and $o(n)$)? We consider convergence in the first k moments of the graph spectrum (equivalent to the numbers of closed k-walks) appropriately normalized. We introduce a simple, easy to sample, random graph model that captures the limiting spectra of many sequences of interest, including the sequence of hypercube graphs. The Random Overlapping Communities (ROC) model is specified by a distribution on pairs (s, q), $s \in \mathbb{Z}_+, q \in (0, 1]$. A graph on n vertices with average degree d is generated by repeatedly picking pairs (s, q) from the distribution, adding an Erdős-Rényi random graph of edge density q on a subset of vertices chosen by including each vertex with probability s/n, and repeating this process so that the expected degree is d. Our proof of convergence to a ROC random graph is based on the Stieltjes moment condition. We also show that the model is an effective approximation for individual graphs. For almost all possible triangle-to-edge and four-cycle-to-edge ratios, there exists a pair (s, q) such that the ROC model with this single community type produces graphs with both desired ratios, a property that cannot be achieved by stochastic block models of bounded description size. Moreover, ROC graphs exhibit an inverse relationship between degree and clustering coefficient, a characteristic of many real-world networks.

Contents

[*]Georgia Tech, spetti@gatech.edu. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650044.

[†]Georgia Tech, vempala@gatech.edu. Both authors were supported in part by NSF awards CCF-1563838 and CCF-1717349.

1 Introduction

What is a good summary of a very large graph? Besides simple statistics like its size and edge density, one would like to know the chance of finding small subgraphs (e.g., triangles), to estimate global properties (e.g., the size of the minimum or maximum cut), and to be able to produce a smaller graph of desired size with similar properties as the original. One approach is to define a random graph model, a simple description of a probability distribution over all graphs, such that a graph drawn from the model will likely have similar properties as the graph of interest. Since the introduction of the Erdős-Rényi random graph $G_{n,p}$ [13, 14] developing and analyzing random graphs has become a rich field with many interesting emergent phenomena.

A more powerful approach to graph approximation is via Szemerédi's regularity lemma [39], which guarantees the existence of a partition the vertices of a graph into a small number of blocks such that the distribution of edges within a block and between most pairs of blocks resembles a random graph of prescribed density. For any dense graph (with $\Omega(n^2)$ edges), the number of partition classes (and hence the total size of the description of the approximation) required to produce an approximation with absolute error $\epsilon > 0$ in the cut norm is only a function of ϵ, independent of the size of the original graph. Frieze and Kannan's weak regularity theorem [16] is a simpler version of Szemerédi's fundamental theorem which gives a weaker approximation (additive ϵ approximation in the cut norm), but produces a much smaller description, leading to efficient approximation algorithms. The regularity lemma's consequences are striking — one can approximate the homomorphism density of any fixed size graph (from the left or right), the size of any cut to within additive error; the partition itself can be constructed algorithmically and is easy to sample. This has lead to effective Stochastic Block Models for large dense graphs. The closely related theory of graph limits shows that any sequence of dense graphs has a convergent subsequence, whose limit captures the limit of homomorphism densities and normalized cuts of the graphs, and is itself a probability distribution over the unit square (called a *graphon*). Moreover, if two graphs are close in the cut metric, then their homomorphism densities are also close. At a qualitative level, these theorems give an essentially complete theory for the approximation of *dense* graphs, where the number of edges is $\Omega(n^2)$.

Such a theory is missing for sparse graphs (with $o(n^2)$ edges). All the approximations for the dense case produce the trivial approximation of the empty graph. While there is an intricately developed theory for bounded-degree graphs that allows one to describe the limits of sequences of such graphs[1], it is not algorithmically tractable, and it does not extend to graph sequences when the degree can grow with the size of the graph. Moreover, as we will presently see, the known objects for approximating dense graphs (block models, regularity decompositions, graphons) are inherently unable to approximate sparse graphs. The main motivation for this paper is to understand what properties of sparse graphs (resp. graph sequences) can be succinctly approximated and to provide a model (limit object) for them. Existing theories are limited in what they can achieve for families of graphs which are neither dense nor bounded-degree. In particular, they seem unable to answer the following representative question [21]:

What is the limit of the sequence of hypercube graphs?

We focus on approximating the simple cycle and closed walk counts of sparse graphs and graph sequences appropriately normalized. These counts encode information about the local structure of the graph and are related to its spectral properties; the number of closed k-walks in a graph is equal to the k^{th} moment of the graph's eigenspectrum. For dense graphs, stochastic block models and graphons approximate both homomorphism densities and cut norm. However the standard cut norm is not useful for sparse graphs as the norm tends to zero. Moreover, natural normalizations do not seem to work either, i.e., they either go to zero or distinguish hypercubes of different sizes.

Another reason we focus on cycle and walk counts rather than cuts is that approximating local structure is of particular interest in practice. A widely-used technique for inferring the structure and function of a real-world graph is to observe overrepresented motifs, i.e., small subgraphs that appear frequently. Recent work describes the overrepresented motifs of a variety of graphs including

[1] One can approximate a bounded-degree graph as a distribution over local neighborhood structures, i.e., the probability that the r-neighborhood of a vertex is a particular graph. For any r, this is a finite description and it captures homomorphism densities and appears as a limit of a bounded-degree graph sequence.

transcription regulation graphs, protein-protein interaction graphs, the rat visual cortex, ecological food webs, and the internet (WWW), [43, 5, 35, 23]. The type of overrepresented motifs has been shown to be correlated with the graph's function [23]. A model that produces graphs with high motif counts is necessary for approximating graphs whose function depends on the abundance of a particular motif.

Limitations of previous approaches for capturing the cycle and walk counts. Previous approaches do not provide a meaningful way to approximate the small cycle counts of sparse graphs. A regularity style partition or stochastic block model inherently cannot approximate the number of triangles unless the rank of the block model grows nearly linearly with the size of the graph as shown in the following simple observation.

Proposition 1.1. *Let M be a symmetric matrix with entries in $[0,1]$ such that each row sum is at most d. Then the expected number of simple k-cycles in a graph obtained by sampling M is at most $d^k rank(M)$.*

The proposition follows by observing that the expected number of simple k-cycles is at most the trace of M^k. In particular, any rank r approximation of the d-dimensional hypercube where each vertex has degree $O(d)$ has fewer than $O(rd^4)$ simple four-cycles, whereas the hypercube has $2^d d^2$ of them.

The local neighborhood distribution approach is hopeless for this setting since the degree is not bounded and therefore there are infinitely many r-neighborhoods [7]. Other methods designed for the sparse but not bounded degree setting do not produce a satisfactory limit object for the sequence of hypercubes. The theory of L^p graphons generalizes the graphon to a range of sparse settings [9]. While the L^p graphon gives approximations for a generalized notion of cut metric for sparse graphs, graphs sampled from the L^p graphon limit of the sequence may have very different normalized subgraph counts than the sequence (i.e. no "counting lemma" is possible). Frenkel redefined homomorphism density with a different normalization based on the size of the subgraph, but this notion does not help distinguish the limiting number of non-tree structures for sequences of graphs with degree tending to infinity [15]. The recently developed notion of *graphex* [40, 8] is the limit object for sequences of sampling convergent graphs. However, any sequence of nearly d-regular graphs with $d = o(n)$ is sampling convergent since the sampled object according to this notion is a set of isolated edges with high probability. Therefore the graphex cannot distinguish between different graph sequences that are nearly regular.

Another natural approach to constructing a graph with high simple cycle density is to repeatedly add simple cycles on a randomly chosen subset of vertices. However, this process yields low cycle 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 62.) In [27] 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).

Normalizing closed walk and cycle counts. In order to meaningfully compare the closed walk counts and cycle counts between graphs of different sizes, it is necessary to normalize the counts. For dense graphs, homomorphism density of a subgraph H in a graph on n vertices is the number of copies of a subgraph H divided by $n^{|v(H)|}$. This normalization is natural because it gives the probability H is present on a random subset of vertices. For the sparse case, this normalization causes the homomorphism density of all subgraphs tend to zero, and so we must define a different normalization.

When considering a sequence of graphs, we can find a proper normalization of the closed walk counts by looking at the rate of growth of the counts. A graph that locally looks like a d-ary tree has approximately $d^{k/2}$ closed k-walks at each vertex for k even. Therefore the appropriate normalization of the closed k-walk counts for a sequence of such graphs is $nd^{k/2}$. We will see in Section 3.1 that this normalization is also natural for the sequence of hypercubes. A sequence of sparse graphs in which each vertex's local neighborhood is dense (e.g., a collection of d cliques of size d), the appropriate normalization for the walk counts is nd^{k-1}. We define the *sparsity exponent* of a sequence to measure the rate of growth of the number of closed k-walks in the sequence. Let $W_k(G)$ be the number of simple cycles of length k in a graph G. [2]

Definition 1. *For $0 < \alpha \leq 1$, we define the α normalized closed k-walk count as*

$$W_k(G, \alpha) = \frac{W_k(G)}{nd^{1+\alpha(k-2)}}$$

where $n = |V(G)|$ and d is the average degree of G.

Definition 2 (sparsity exponents)**.** *Let (G_i) be a sequence of graphs. Let*

$$\alpha = \inf_{b \in [1/2,1]} \left\{ b \mid \lim_{i \to \infty} W_j(G_i, b) \text{ exists for all } j \right\}$$

be the sparsity exponent of the sequence. For $k \geq 3$, let

$$\alpha_k = \inf_{b \in [1/2,1]} \left\{ b \mid \lim_{i \to \infty} W_j(G_i, b) \text{ exists for all } j \leq k \right\}$$

be the k-sparsity exponent of the sequence.

We define the minimum of the sparsity exponent to be $1/2$ because all d-regular graphs have at least $Cat_{k/2} nd^{k/2}$ closed k-walks obtained from tracing trees. For two sequences of graphs with matching degrees, a higher sparsity exponent indicates denser local neighborhoods and therefore more closed walks.

While the sparsity exponent gives a natural way to normalize the closed walk counts for a sequence of sparse graphs, it does not help determine the appropriate normalization factor for approximating an individual graph (not contextualized in a sequence). When approximating an individual graph, we instead choose to focus on the number of simple k-cycles denoted $C_k(G)$ and normalize by the number of edges in the graph. Throughout this paper, for convenience we refer to a simple k-cycle as a k-cycle. For example, under this convention each triangle is counted 6 times because there are 6 closed walks that traverse a triangle.

As described in the previous section, approximating $W_k(G)$ and $C_k(G)$ for sparse graphs is already out-of-reach for known methods that work well in the dense and bounded-degree settings. The main contribution of this paper is the following model which can approximate the normalized closed walk and cycle counts for a large class of graphs.

[2] Denoting the number of closed k-walks by $W_k(G)$, and the eigenvalues of the adjacency matrix of G as $\lambda_1(G) \geq \lambda_2(G), \cdots \geq \lambda_n(G)$, we have $W_k(G) = \sum_{i=1}^{n} \lambda_i^k$.

The Random Overlapping Communities Model. We introduce a simple generalization of the Erdős-Rényi model that can approximate the normalized cycle and walk counts of a wide range of sparse graphs. The *Random Overlapping Communities (ROC)* model generates graphs that are the union of many relatively dense random communities. A *community* is an instance of an Erdős-Rényi graph $G_{s,q}$ (or a bipartite Erdős-Rényi graph $G_{s/2,s/2,q}$) on a set of s randomly chosen vertices. A ROC graph is the union of many randomly selected communities that overlap, so every vertex is a member of multiple communities. The number, size and density of communities are drawn from a distribution. Figure 1 illustrates this construction.

An instantiation of the ROC model is given by a distribution $\mathcal{D}$ on triples (s, q, b) where s is an integer, $0 \le q \le 1$ and $b < s$ is an integer indicating bipartiteness. A graph of a desired size n and expected degree d is generated by repeatedly selecting a triple (s, q, b) from the distribution $\mathcal{D}$ and picking each vertex with probability s/n and adding a random graph of edge density q in the subgraph. We refer to each such structure as a community. If $b > 0$, then the subset of s vertices is partitioned into two subsets of expected size $b, s - b$ and edges are added only between these subsets. For this paper we will set $b = 0$ or $b = s/2$. See Section 4.1 for a formal definition of the model. We note that the ROC model can be viewed as a generalization of the Erdős-Rényi model (with $s = n$ and $q = d/n$) and maintains the property that it is easy to sample given its defining parameters.

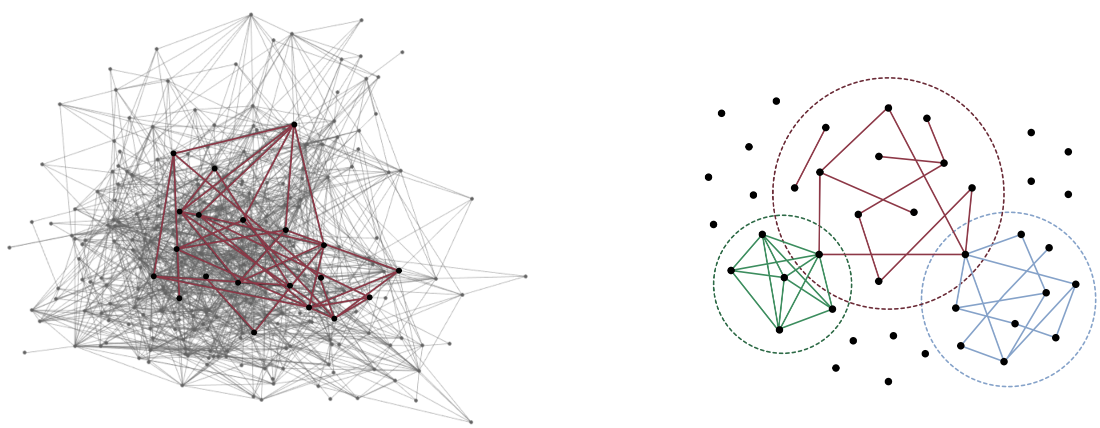

Figure 1: Left: in each step of the construction of a ROC(n, d, s, q) graph, an instance of $G_{s,q}$ is added on a set of s randomly selected vertices. Right: three communities of a ROC graph.

Organization. The paper has two main objectives: to show that ROC is an effective approximation for individual sparse graphs (Section 2) and to develop a theory of sparse graph limits in which the ROC model is a natural limit object (Sections 3 to 5). These two parts are self-contained and may be read independently. We end with a discussion of limitations of the model, possible extensions, and open questions (Section 6). In the remainder of this section we summarize the results.

1.1 ROC for approximating a single graph

In Section 2, we show that the ROC model can approximate the triangle-to-edge and four-cycle-to-edge ratios of a graph, and can be tuned to exhibit high clustering coefficient (the probability two randomly selected neighbor of a random vertex are adjacent). Since these properties are of interest in practice, the model may be of use in real-world contexts. In addition, we introduce a variant of the model that produces graphs with varied degree distributions. For a comparison of the ROC model to existing models used in practice see Section 2.4.

First, we show that for almost all triangle-to-edge and four-cycle-to-edge ratios arising from some graph, there exists a single community size s density q such that the ROC model produces 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.

Theorem 1.

1. *Let H be a graph and let $c_i = C_k(H)/|E(H)|$ for $i = 3, 4$. Then $c_3(c_3/2 - 1) \leq c_4$.*
2. *For any c_3 and c_4 such that $c_3^2 \leq 2c_4$, and $d = o(n^{1/3})$, the random graph $G \sim ROC\,(n, d, \mathcal{D})$ where $\mathcal{D}$ is the distribution with support one on $s = \frac{2c_4^2}{c_3^3}$ and $q = \frac{c_3^2}{2c_4}$ has*

$$\lim_{n\to\infty} \frac{2\,\mathsf{E}[C_3(G)]}{nd} = c_3 \quad \text{and} \quad \lim_{n\to\infty} \frac{2\,\mathsf{E}[C_4(G)]}{nd} = c_4.$$

Theorem 8 gives conditions for determining when it is possible to construct a ROC family that matches a vector of k-cycle-to-edge ratios. These conditions are related to the conditions for determining when the ROC model is the limit object for a sequence of graphs.

Modeling the clustering coefficient of real-world graphs. In Theorem 9, we prove the average clustering coefficient of a ROC graph (with one community type) 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 10 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 networks [38, 22, 24, 2].

Diverse degree distributions and the DROC model. We also introduce an extension of our model which produces graphs that match a target degree distribution in expectation. The extension uses 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 $\frac{d_i d_j}{\sum_{i=1}^n d_i}$, yielding a graph where the expected degree of vertex v_i is d_i [11]. 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.

1.2 ROC as a limit object for sparse graph sequences

We show that the ROC model is a limit for several interesting sequences of graphs and a give a characterization of sequences which are the limits of ROC. To state our results, we first define the convergence of sparse graph sequences and their limits. We consider convergence first for each k and then for all positive integers k, referring to the latter as full convergence. In Section 3, we compute the limits for the hypercube sequence, the rook's graph sequence (a family of strongly regular graphs) as well as for Erdős-Rényi random graphs.

Definition 3 (k-convergent)**.** *Let (G_i) be a sequence of graphs with k-sparsity exponent α_k. The sequence (G_i) is k-convergent if $\lim_{i\to\infty} W_j(G_i, \alpha_k)$ exists for all $j \le k$. We let $w_j = \lim_{i\to\infty} W_j(G_i, \alpha_k)$ and say $(w_3, w_4, \dots, w_k)$ is the k-limit of the graph sequence (G_i).*

Definition 4 (fully convergent)**.** *Let (G_i) be a sequence of graphs with sparsity exponent α. We say the sequence is fully convergent if $\lim_{i\to\infty} W_j(G_i, \alpha)$ exists for all j. We let $w_j = \lim_{i\to\infty} W_j(G_i, \alpha)$ and say $(w_3, w_4, \dots)$ is the limit of the graph sequence (G_i).*

Informally, we say that a ROC family (a distribution on triples) achieves the limit of a convergent sequence of graphs (G_i) if the normalized expected number of walks in a graph drawn from the ROC family matches the limit of (G_i). We consider the ROC family that achieves the limit to be a limit object of (G_i). Since a particular limit may be achieved by many ROC families, the limit object for a sequence is not unique.

We now formalize the notion of a ROC family achieving a vector of normalized counts as its limit. We use *achievable* to describe when a ROC family realizes a k-limit, *fully achievable* to describe when a ROC family realizes a limit, and *totally k-achievable* to describe the weaker notion that any subsequence of a limit is achievable by a ROC family.

Definition 5 (k-achievable, totally k-achievable, fully achievable)**.**

1. *The k-limit $(w_3, w_4, \dots w_k)$ of a sequence of graphs with sparsity exponent α is* k-achievable by ROC *if there exists a ROC family $\mathcal{D}$ such that for all $3 \le j \le k$, when $d \to \infty$ and $d = o\left(n^{1/((1-a)k+2a-1)}\right)$*
$$\lim_{n\to\infty} \frac{\mathsf{E}[W_j\left(ROC(n, d, \mathcal{D})\right]}{nd^{1+\alpha(j-2)}} = w_j.$$

2. *The limit of a sequence of graphs* totally k-achievable by ROC *if every k-limit of the sequence is achievable (possibly with a different choice for each k).*

3. *The limit $(w_3, w_4, \dots)$ of a sequence of graphs with sparsity exponent α is* fully achievable by ROC *if there exists a ROC family $\mathcal{D}$ such that for all $j \ge 3$, when $d \to \infty$ and $o\left(n_i^{\varepsilon}\right)$ for all ε if $a < 1$ and $d_i = o(n_i)$ for $a = 1$*
$$\lim_{n\to\infty} \frac{\mathsf{E}[W_j\left(ROC(n, d, \mathcal{D})\right)]}{nd^{1+\alpha(j-2)}} = w_j.$$

Roughly speaking, the degree upper bounds ensure that the overwhelming majority of simple cycles are contained entirely in single communities. In Theorem 34 we show that the probability that the normalized closed walk counts of $G \sim ROC(n, d, \mathcal{D})$ deviate from the family's limit vanishes as $d \to \infty$. Moreover Corollary 35 gives conditions on n_i and d_i which guarantee that a sequence (G_i) with $G_i \sim ROC(n_i, d_i, \mathcal{D})$ almost surely converges to limit vector achieved by the family.

Results for convergent sequences. We begin with the limit of the sequence of hypercube graphs, answering the question raised by [21].

Theorem 2. *The limit of the sequence of hypercube graphs is totally k-achievable by ROC.*

This theorem generalizes to sequences of Hamming cubes and Cayley graphs of $(\mathbb{Z} \mod k\mathbb{Z})^d$ (Corollary 54). These sequences have the same limit as the hypercube sequence and therefore are achieved by the same ROC family.

The next theorem is about a sequence of strongly regular graphs called rook's graphs (the Cartesian product of two complete graphs, see Lemma 21).

Theorem 3. *The limit of the sequence of rook's graphs is fully achievable by ROC.*

We also discuss the convergence of sequences of Erdős-Rényi random graphs (Lemma 24), demonstrating the limits of some sequences cannot be achieved exactly by ROC familes, but they can be approximated to arbitrarily small error.

Achievability and the Stieltjes condition. A limit vector L is achieved by a ROC family when the normalized expected walk counts of a graph sampled from the ROC family match L up to terms that vanish as the size of the sampled graph grows. The number of closed walks in a ROC graph is related to its simple cycles counts, and the expected simple cycle counts are the moments of a distribution determined by the ROC parameters. Therefore, determining which vectors can be achieved by a ROC family is closely related to the Stieltjes' moment problem: given a sequence whether there exists a discrete distribution with positive support with that moment sequence? This is the classical Stieltjes moment problem, whose solution is characterized by the definition below (see Lemma 43).

Definition 6 (Stietljes conditions)**.** *The Hankel matrices of a sequence μ are*

$$H_{2s}^{(0)} = \begin{pmatrix} \mu_0 & \mu_1 & \cdots & \mu_s \\ \mu_1 & & & \\ \vdots & & \ddots & \vdots \\ \mu_s & & \cdots & \mu_{2s} \end{pmatrix} \quad \text{and} \quad H_{2s+1}^{(1)} = \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_{s+1} \\ \mu_2 & & & \\ \vdots & & \ddots & \vdots \\ \mu_{s+1} & & \cdots & \mu_{2s+1} \end{pmatrix}.$$

1. *The vector $\mu = (\mu_0, \mu_1, \dots \mu_n)$ satisfies the Stieltjes condition if*

$$det\left(H_{2s}^{(0)}\right) \geq 0 \text{ for all } 0 \leq 2s \leq n \quad \text{and} \quad det\left(H_{2s+1}^{(1)}\right) \geq 0 \text{ for all } 1 \leq 2s+1 \leq n,$$

and for k the smallest integer such that $det\left(H_{2k}^{(0)}\right) = 0$ or $det\left(H_{2k+1}^{(1)}\right) = 0$,

$$det\left(H_{2i}^{(1)}\right) = 0 \quad \text{and} \quad det\left(H_{2i+1}^{(1)}\right) = 0 \quad \text{for all} \quad k \leq i \leq n.$$

2. *The infinite vector $\mu = (\mu_0, \mu_1, \dots)$ satisfies the full Stieltjes condition if the above statements hold for all n.*

In ROC families that produce sequences of graphs with sparsity exponent greater than 1/2, the counts of simple cycles dominate the total closed walk counts. Therefore, a limit is achievable when it is possible to construct a ROC family with normalized simple cycle counts that match the desired normalized closed walk counts. The cycle counts are dominated by the cycles contained entirely in one community. Every community contributes even cycles, but only the non-bipartite communities contribute to the odd cycle counts. In the following theorems, the parameters s_i and t_i count the number of simple i-cycles in non-bipartite and bipartite communities respectively, and the parameter γ indicates the expected fraction of communities that are non-bipartite.

Theorem 4 (achievability with sparsity exponent $> 1/2$)**.** *A limit vector $(w_3, w_4, \dots w_k)$ is achievable by ROC with sparsity exponent greater than $1/2$ if and only if there exists $\gamma \in [0,1]$, $s_0, s_1, \dots s_k, t_0, t_2, \dots t_{2\lfloor \frac{k}{2} \rfloor} \in \mathbb{R}^+$, $s_2, t_2 \leq 1$ such that $(s_0, s_1, s_2, \dots s_k)$ and $(t_0, t_2, \dots t_{2\lfloor \frac{k}{2} \rfloor})$ satisfy the Stieltjes condition and for $3 \leq j \leq k$*

$$w_j = \begin{cases} \gamma s_j & j \text{ odd} \\ \gamma s_j + (1-\gamma) t_j & j \text{ even.} \end{cases}$$

Approximating sequences with sparsity exponent 1/2 is more complicated because the number of simple cycles can be of the same order as the number of closed walks that are not simple cycles in ROC families that produce graph sequences with sparsity exponent 1/2. In Theorem 27 we prove that the polynomial T given in Definition 8 describes the relationship between simple cycle counts and closed walks counts in ROC graphs. Moreover we show that T describes the relationship between simple cycle counts and closed walk counts in *locally regular* graphs in which each vertex is in the same number of cycles. A limit with sparsity exponent 1/2 is achievable when it is possible to construct a ROC family with normalized simple cycle counts that match the inverse of this polynomial T applied to the desired normalized closed walk counts.

Theorem 5 (achievability with sparsity exponent 1/2)**.** *Let $T((c_3, c_4, \dots c_k)) = (w_3, w_4, \dots w_k)$ be the transformation of a vector given in Definition 8. The limit vector $(w_3, w_4, \dots w_k)$ is achievable by ROC with sparsity exponent $1/2$ if and only if there exists $\gamma \in [0,1]$, $s_0, s_1, s_2, \dots s_k, t_0, t_2, \dots t_{2\lfloor \frac{k}{2} \rfloor} \in \mathbb{R}^+$, $s_2, t_2 \leq 1$ such that $(s_0, s_1, s_2, \dots s_k)$ and $(t_0, t_2, \dots t_{2\lfloor \frac{k}{2} \rfloor})$ satisfy the Stieltjes condition and for $3 \leq j \leq k$*

$$c_j = \begin{cases} \gamma s_j & j \text{ odd} \\ \gamma s_j + (1-\gamma) t_j & j \text{ even.} \end{cases}$$

The analogous theorems for the full achievability of limits by ROC require the full Stieltjes condition. See Theorems 40 and 41 in Section 4.3.2. The Stieltjes condition also determines when a vector of k-cycle-to-edge ratios can be matched by a ROC family, demonstrating the relevance of our method for different normalizations (Theorem 8). All 4-limits can be achieved by a ROC model; however not all k-limits can be achieved. In Section 6.1 we give an example of a graph sequences with a 6-limit that cannot be achieved by a ROC family.

Theorem 6. *The limit (w_3, w_4) of any convergent sequence of graphs with increasing degree is achieved by a ROC family.*

We have already seen that any realizable triangle and four-cycle count normalized by the number of edges can be approximated by a ROC model with one community type.

2 Approximating a graph with a ROC

In this section we consider the utility of the ROC model for approximating individual graphs. First we show that almost all pairs of triangle-to-edge and four-cycle-to-edge ratio can be approximated with a ROC graph with one community type and give conditions for when a vector of k-cycle-to-edge ratios can be achieved by ROC generally. (In Appendix B, we analyze the connectivity of ROC graphs.) Then, we shift our focus to modeling real-world graphs with ROC. In Section 2.2 we show that ROC graphs exhibit high clustering coefficient and an inverse relationship between clustering coefficient and degree, a phenomena observed in real world networks. In Section 2.3 we introduce an extension of the ROC model that produces graphs with varied degree distributions. Finally we end the section by comparing the ROC model to existing models used in practice (Section 2.4).

Often in this section, we focus on the special case of the ROC model when the distribution of communities $\mathcal{D}$ is taken to be a single community s and density q. When this is clear from context, we denote the model $ROC(n, d, s, q)$. In terms of the formal parameterization of the ROC model given in Section 4.1, $a = 0$ and μ is the the distribution with support one on $m_i = s$, $q_i = q$ and $\beta_i = 0$.

2.1 The k-cycle-to-edge ratios of ROC graphs

In this section we prove Theorem 1, which states that most triangle-to-edge and four-cycle-to-edge ratios can be approximated simultaneously by the ROC model on one community. Then we prove Theorem 8, which describes more generally when it is possible to match all j-cycle-to-edge ratios up to some k with the ROC model.

To begin we consider the ROC model with all communities of size s and density q. The following lemma describes the k-cycle-to-edge ratios of ROC graphs in this setting. The lemma is a special case of Corollary 33.

Lemma 7. *Let $G \sim ROC(n, d, s, q)$. Then*

$$R_k = \lim_{n\to\infty} \frac{2\,\mathsf{E}[C_k(G)]}{nd} = 2s^{k-2}q^{k-1} \textit{ for } d = o(n^{1/(k-1)}).$$

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 2 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 $R_4 \in \{3, 10\}$.

We apply Lemma 7 to prove Theorem 1, which states that any non-zero triangle and four-cycle ratios satisfying $c_3^2 \leq 2c_4$ can be approximated with the ROC model. For every graph with triangle and four-cycle ratios in the narrow range $c_3(c_3/2-1) \leq c_4 \leq c_3^2/2$, there exists a ROC construction that matches c_3 and can approximate c_4 by $c_3^2/4$, i.e., up to an additive error $c_3/8$ (or multiplicative error of at most $1/(c_3/2 - 1)$ which goes to zero as c_3 increases).

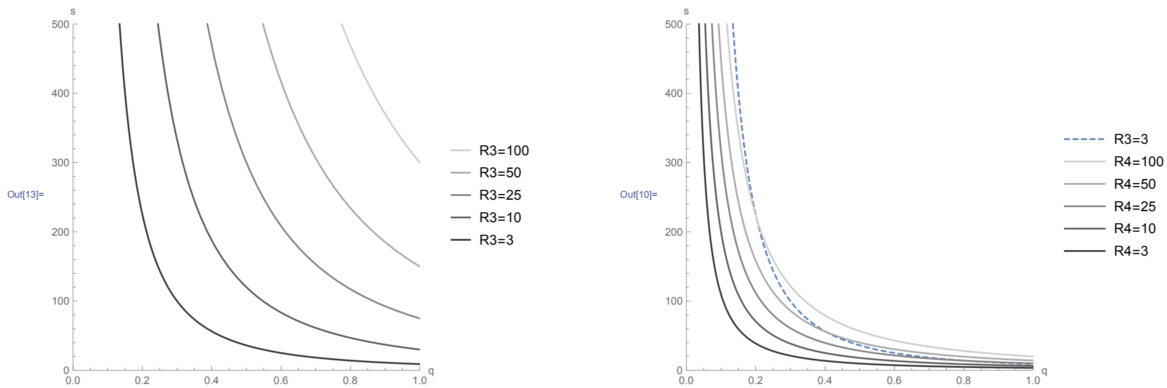

Figure 2: Left: A wide range of s and q yield the same R_3 and R_4 ratio (left and right respectively).

Proof. (of Theorem 1.) (1) For clarity of this proof we refer to the number triangle and four-cycle structures (not counted as walks). Under this convention, the number of triangles is $T_3 = C_3(H)/6$ and the number of four-cycles is $T_4 = C_4(H)/8$. Note $T_3 = |E(H)|c_3/6$ and $T_4 = c_4|E(H)|/8$. For each edge in H, let t_e be the number of triangles containing e, so $\sum_{e\in E(H)} t_e = 3T_3 = c_3|E(H)|/2$. 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 $T_4 \geq \frac{1}{2}\sum_{e\in E(H)}\binom{t_e}{2}$. This expression is minimized when all t_e are equal. We therefore obtain

$$\frac{c_4|E(H)|}{8} = T_4 \geq \frac{|E(H)|}{2}\binom{c_3/2}{2} = \frac{c_3(c_3/2-1)|E(H)|}{8}.$$

It follows that $\frac{c_3(c_3/2-1)}{c_4} \leq 1$.

(2) Since the hypothesis guarantees $q \leq 1$, applying Lemma 7 to $G \sim ROC\left(n, d, \frac{2c_4^2}{c_3^3}, \frac{c_3^2}{2c_4}\right)$ implies the desired statements. ∎

By increasing the support size of the distribution over communities, it is possible to achieve a wider range of k-cycle-to-edge-ratios. The following theorem shows that the condition for determining whether a vector of k-cycle-to-edge ratios of a graph can be matched by a ROC family is closely related to determining if a limit of a sequence of graphs is achievable by a ROC family.

Theorem 8. *There exists a ROC family such that for $G \sim ROC(n, d, \mathcal{D})$ with $d = o\left(n^{1/(k-1)}\right)$*

$$\lim_{n\to\infty} \frac{2\,\mathsf{E}[C_j(G)]}{nd} = c_j \quad \text{for } 3 \leq j \leq k$$

if and only if there exists $\gamma \in [0,1]$, $s_0, s_1, s_2, \ldots s_k, t_0, t_2, \ldots t_{2\lfloor \frac{k}{2}\rfloor} \in \mathbb{R}^+$, $s_2, t_2 \leq 1$ such that $(s_0, s_1, s_2, \ldots s_k)$ and $(t_0, t_2, \ldots t_{2\lfloor \frac{k}{2}\rfloor})$ satisfy the Stieltjes condition and for $3 \leq j \leq k$

$$c_j/2 = \begin{cases} \gamma s_j & j \text{ odd} \\ \gamma s_j + (1-\gamma)t_j & j \text{ even.} \end{cases}$$

The proof of the above theorem is a slight modification of the proofs of the main limit achievability theorems (Theorem 5, Theorem 4), and so we give the proof in Section 4.3.2. Later we show

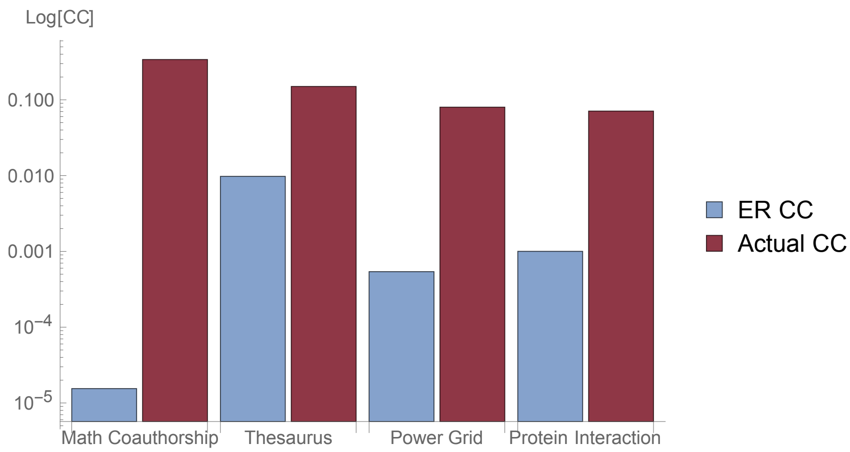

Figure 3: The clustering coefficient in real world graphs is much greater than that of an E-R random graph of the same density. Data from Table 3.1 of [25].

that a ROC family that achieves a normalized closed walk count limit is parameterized so that the community sizes grow with d^a for some constant $a \in [1/2, 1]$. In the proof of the above theorem we show that ROC family that approximates a vector of k-cycle-to-edge ratios will have constant community sizes.

2.2 Approximating clustering coefficient

Closely related to the density of triangles is the clustering coefficient at a vertex v, 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. Figure 3 illustrates the markedly high clustering coefficients of real-world graphs as compared with Erdős-Rényi (E-R) graphs of the same density. We show that the ROC model can be tuned to produce graphs with a variety of clustering coefficients at any density. The proofs in this section are quite technical and left to Section 2.2.1.

Theorem 9 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 13 of Section 2.2.1, and bounds in a more general setting are given by (4).

Theorem 9. *Let $C(v)$ denote the clustering coefficient of a vertex v with degree at least 2 in a graph drawn from $ROC(n, d, s, q)$ with $d = o(\sqrt{n})$, $d < (s-1)qe^{sq}$, $d = \omega(sq \log \frac{nd}{s})$, $s^2 q = \omega(1)$, and $sq = o(d)$. Then*

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

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

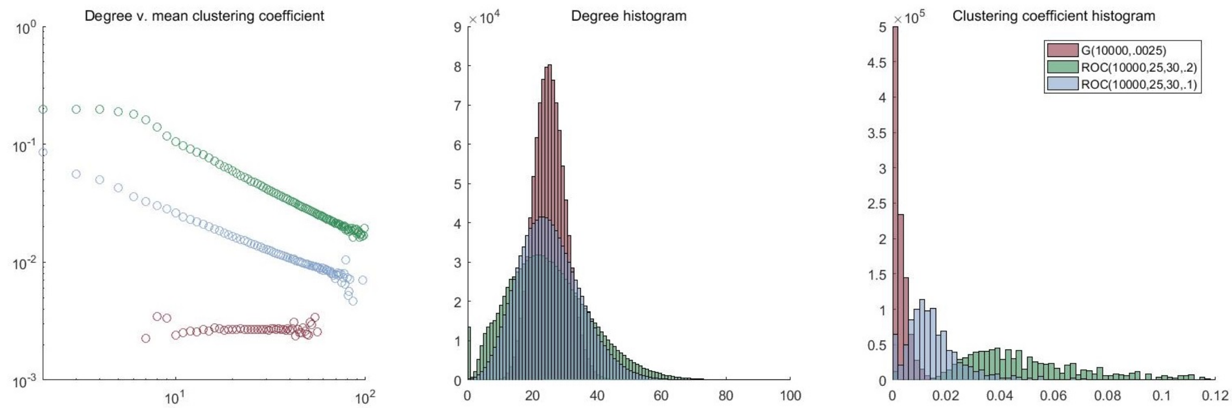

Figure 4: A comparison of the degree distributions and clustering coefficients of 100 graphs with average degree 25 drawn from each $G_{10000,0.0025}$, $ROC(10000, 25, 30, 0.2)$, and $ROC(10000, 25, 30, 0.1)$. The mean clustering coefficients are 0.00270, 0.06266, and 0.01595 respectively.

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

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

2.2.1 Clustering coefficient proofs

Remark 11. *Theorem 9 gives bounds on the 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 9 that $d = \omega(sq \log \frac{nd}{s})$, $d < (s-1)qe^{sq}$, and $s = \omega(1)$, Lemma 12 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 9 and supporting Lemma 13. In addition we divide by $deg(v)^2$ rather than by $deg(v)(deg(v) - 1)$ in the computation of the clustering coefficient since this modification only affects the computations up to a factor of $(1 + o(1))$.*

Lemma 12. *If $d = \omega(sq \log \frac{nd}{s})$, $s = \omega(1)$, $s = o(n)$, and $d < (s-1)qe^{sq}$, then a graph from $ROC(n, d, s, q)$ a.a.s. has no vertices of degree less than 2.*

Proof. Theorem 63 implies there are no isolated vertices a.a.s. We begin by computing the probability a vertex has degree one.

$$
\begin{aligned}
\mathsf{Pr}[deg(v)=1] &= \sum_{i=1}^{\frac{nd}{s^2q}} \mathsf{Pr}[v \text{ is in } i \text{ communities}] q(1-q)^{si-1} \\
&= \sum_{i=1}^{\frac{nd}{s^2q}} \binom{\frac{nd}{s(s-1)q}}{i} \left(\frac{s}{n}\right)^i \left(1-\frac{s}{n}\right)^{\frac{nd}{s(s-1)q}-i} q(1-q)^{si-1} \\
&\leq (1+o(1)) \sum_{i=1}^{\frac{nd}{s^2q}} \left(\frac{nd}{s(s-1)q}\right)^i \left(\frac{s}{n}\right)^i e^{-\frac{d}{sq}+\frac{si}{n}} q e^{-qsi+q} \\
&= (1+o(1))\, q e^{-\frac{d}{sq}} \sum_{i=1}^{\frac{nd}{s^2q}} \left(\frac{de^{-sq}}{(s-1)q}\right)^i \\
&= O\left(\frac{de^{-sq-\frac{d}{sq}}}{s}\right)
\end{aligned}
$$

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

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

□

Lemma 13. *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*

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

Proof. For ease of notation, we ignore factors of $(1+o(1))$ throughout as described in Remark 11. 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 $X_1,\dots X_k$ be random variables representing the degree of v in each of the communities, $X_i \sim Bin(s,q)$. We have

$$
\begin{aligned}
\mathsf{E}[C(v)|\ v \text{ in } k \text{ communities }] &= \mathsf{E}\left[\frac{\sum_{i=1}^k X_i(X_i-1)q}{\left(\sum_{i=1}^k X_i\right)^2}\right] && (1) \\
&= qk\,\mathsf{E}\left[\frac{X_1(X_1-1)}{(sq(k-1)+X_1)^2}\right] \\
&= qk\,\mathsf{E}\left[\frac{X_1^2}{(sq(k-1)+X_1)^2}\right] - qk\,\mathsf{E}\left[\frac{X_1}{(sq(k-1)+X_1)^2}\right].
\end{aligned}
$$

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

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

and

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

Substituting in these values into (1), we obtain

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

Let M be the number of communities a vertex is in, so $M \sim Bin\left(\frac{nd}{s^2q}, \frac{s}{n}\right)$. It follows

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

□

The proof of Theorem 9, relies on the follow two lemmas regarding expectation of binomial random variables.

Lemma 14. *Let $X \sim Bin(n,p)$. Then*

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

Proof. Observe

$$\begin{aligned}
\mathsf{E}\left[\frac{1}{X+1} \mid X \geq 1\right] &= \sum_{i=1}^{n} \binom{n}{i} \frac{p^i(1-p)^{n-i}}{i+1} \\
&= \frac{1}{p(n+1)} \sum_{i=1}^{n} \binom{n+1}{i+1} p^{i+1}(1-p)^{n-i} \\
&= \frac{1-(1-p)^{n+1}-(n+1)p(1-p)^n}{p(n+1)}.
\end{aligned}$$

Similarly

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

□

Lemma 15. *Let $X \sim Bin(n,p)$. Then*

$$\mathsf{E}\left[\frac{1}{X} \mid X \geq 1\right] \leq \frac{1}{p(n+1)}\left(1+\frac{3}{p(n+2)}\right).$$

Proof. Note that when $X \geq 1$,

$$\frac{1}{X} \leq \frac{1}{X+1} + \frac{3}{(X+1)(X+2)}.$$

By Lemma 14,

$$\mathsf{E}\left[\frac{1}{X+1} \mid X \geq 1\right] \leq \frac{1}{p(n+1)}. \tag{3}$$

We compute

$$\begin{aligned}
\mathsf{E}\left[\frac{1}{(X+1)(X+2)} \mid X \geq 1\right] &= \sum_{i=1}^{n} \frac{\binom{n}{i} p^i(1-p)^{n-i}}{(i+1)(i+2)} \\
&= \frac{1}{p^2(n+2)(n+1)} \sum_{i=1}^{n} \binom{n+2}{i+2} p^{i+2}(1-p)^{n-i} \\
&\leq \frac{1}{p^2(n+2)(n+1)}.
\end{aligned}$$

Taking expectation of (3) gives

$$\mathsf{E}\left[\frac{1}{X} \mid X \geq 1\right] \leq \frac{1}{p(n+1)}\left(1+\frac{3}{p(n+2)}\right).$$

□

Proof. (of Theorem 9.) For ease of notation, we ignore factors of $(1 + o(1))$, as described in Remark 11. It follows from (2) in the proof of Lemma 13 that

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

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

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

$$\mathsf{E}[C(v)] = \sum_{k=1}^{\frac{nd}{s^2 q}} \mathsf{Pr}[M=k]\, \mathsf{E}[C(v)|M=k]$$

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

Applying Lemmas 14 and 15 to the lower and upper bounds respectively, we obtain

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

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

$$(1+o(1))\,\frac{(s-1)q^2}{d}\left(1-\frac{nd}{s(s-1)q}e^{-d/((s-1)q)}\right) \leq \mathsf{E}[C(v)] \leq \frac{(s-1)q^2}{d}\left(1+\frac{(s-1)q}{d}\right)(1+o(1))\,. \tag{4}$$

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

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

□

The following lemma will be used in the proof of Theorem 10.

Lemma 16. *The X be a nonnegative integer drawn from the discrete distribution with density proportional to $f(x) = x^{r-x}e^{-ax}$. Let $z = \arg\max f(x)$. Then*

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

Proof. First we observe that f is logconcave:

$$\frac{d^2}{dx^2}\ln f(x) = \frac{d}{dx}(-a + \frac{r}{x} - 1 - \ln x) = -\frac{r}{x^2} - \frac{1}{x}$$

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

$$a + 1 = \frac{r}{x} - \ln x. \tag{5}$$

The maximizer z is very close to

$$\frac{r}{(a+1) + \ln \frac{r}{(a+1)+\ln(r/(a+1))}}, \tag{6}$$

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.,

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

The LHS is

$$\begin{aligned} \left(1+\frac{\delta}{z}\right)^{r-z} z^{-\delta}\left(1+\frac{\delta}{z}\right)^{-\delta} e^{-a\delta} &\leq e^{\delta(\frac{r}{z}-1-a-\ln z)}e^{-\frac{\delta^2}{z}} \\ &\leq e^{-\frac{\delta^2}{z}} \end{aligned}$$

where in the second step we used the optimality condition (5). Thus for $\delta = (1+o(1))\sqrt{z}$, $f(x+\delta) \leq f(x)/e$. By logconcavity (which says that for any x, y and any $\lambda \in [0,1]$, we have $f(\lambda x + (1-\lambda)y) \geq f(x)^\lambda f(y)^{1-\lambda}$) we have

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

for any $t \geq 1$. It follows

$$f(x+t\delta) \leq f(x)/e^t \tag{7}$$

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

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

and so

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

□

Proof. (of Theorem 10). Let M denote the number of communities a vertex v is selected to participate in. We can write

$$\begin{aligned} \mathsf{E}[C(v)|deg(v)=r] &= \sum_{k=\frac{r}{s}}^{r} \mathsf{E}[C(v)|deg(v)=r, M=k]\,\mathsf{Pr}[M=k|deg(v=r] \\ &= \sum_{k=\frac{r}{s}}^{r} \mathsf{E}[C(v)|deg(v)=r, M=k]\,\mathsf{Pr}[deg(v)=r|M=k]\frac{\mathsf{Pr}[M=k]}{\mathsf{Pr}[deg(v)=r]}. \end{aligned}$$

First we compute the expected clustering coefficient of a degree r vertex given that it is k communities:

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

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 $Bin(s,q)$. Therefore,

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

Using this we obtain

$$\begin{aligned}
\mathsf{E}[C(v)|deg(v) = r] &= \frac{\sum_{k=\frac{r}{s}}^{r} \frac{q}{k}\,\mathsf{Pr}[M = k]\,\mathsf{Pr}[deg(v) = r|M = k]}{\sum_{k=\frac{r}{s}}^{r} \mathsf{Pr}[M = k]\,\mathsf{Pr}[deg(v) = r|M = k]} \\
&= (1+o(1))\,q\frac{\sum_{k=\frac{r}{s}}^{r} \frac{1}{k}\cdot\left(\frac{d}{(s-1)qk}\right)^k e^{-\frac{d}{(s-1)q}+\frac{sk}{n}}\left(\frac{skq}{r}\right)^r e^{-qsk+qr}}{\sum_{k=\frac{r}{s}}^{r} \left(\frac{d}{(s-1)qk}\right)^k e^{-\frac{d}{(s-1)q}+\frac{sk}{n}}\left(\frac{skq}{r}\right)^r e^{-qsk+qr}} \\
&= (1+o(1))\,q\frac{\sum_{k=\frac{r}{s}}^{r} \frac{1}{k}\cdot\left(\frac{d}{(s-1)q}\right)^k k^{r-k}e^{-qsk}}{\sum_{k=\frac{r}{s}}^{r} \left(\frac{d}{(s-1)q}\right)^k k^{r-k}e^{-qsk}}. \qquad (8)
\end{aligned}$$

Writing $a = qs - \ln(d/(s-1)q)$, this is

$$q\frac{\sum_{k=\frac{r}{s}}^{r} \frac{1}{k}\cdot k^{r-k}e^{-ak}}{\sum_{k=\frac{r}{s}}^{r} k^{r-k}e^{-ak}}.$$

Therefore (8) is the same as $q\,\mathsf{E}[1/x]$ 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 16, so $z \approx \frac{r}{sq}$. We use Lemma 16 to bound

$$\begin{aligned}
\mathsf{E}\left[\left|\frac{1}{x}-\frac{1}{z}\right|\right] &\le \sum_{t=1}^{\infty}\left(\frac{1}{z}-\frac{1}{z+t\sqrt{z}}\right)e^{-t} + \sum_{t=1}^{\sqrt{z}-1}\left(\frac{1}{z-t\sqrt{z}}-\frac{1}{z}\right)e^{-t} \\
&= \sum_{t=1}^{\infty}\frac{t\sqrt{z}e^{-t}}{z(z+t\sqrt{z})} + \sum_{t=1}^{\sqrt{z}-1}\frac{t\sqrt{z}e^{-t}}{z(z-t\sqrt{z})} \\
&\le \frac{1}{z}\sum_{t=1}^{\infty}\frac{te^{-t}}{\sqrt{z}+1} + \frac{\sqrt{z}}{z}\left(\sum_{t=1}^{\sqrt{z}/3}\frac{3te^{-t}}{2z} + \sum_{t=\sqrt{z}/3}^{\sqrt{z}-1} te^{-t}\right) \\
&= \frac{O(1)}{z\sqrt{z}} + \frac{O(1)}{z\sqrt{z}} + O\left(\frac{\sqrt{z}}{3}e^{-\frac{\sqrt{z}}{3}}\right) = \frac{O(1)}{z\sqrt{z}}.
\end{aligned}$$

Using this and approximating z by $\frac{r}{sq}$, the expectation of x with respect to the density proportional to f can be estimated:

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

as claimed.

□

2.3 Varied degree distributions: the DROC extension

In this section we introduce an extension of our model which produces graphs that match a target degree distribution in expectation. In each iteration a modified Chung-Lu random graph is placed instead of an E-R random graph.

DROC(n, D, s, q).
Input: number of vertices n, target degree sequence $D = t(v_1), \ldots t(v_n)$ with mean d.
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 $\frac{qt(v_i)t(v_j)}{sd}$; if the edge already exists, do nothing.

Theorem 17. *Given a degree distribution D with mean d and $\max_i t(v_i)^2 \le \frac{sd}{q}$, DROC$(n, D, s, q)$ yields a graph where vertex v_i has expected degree $t(v_i)$.*

We require $\max_i t(v_i)^2 \le \frac{sd}{q}$ to ensure that the probability each edge is chosen is at most 1. In the DROC model the number of communities a vertex belongs to is independent of target degree $t(v)$. When $t(v) > \frac{sd}{q}$, if v participates in the average number of communities and is connected to all vertices in each of its communities, it likely will not reach degree $t(v)$. Therefore when s is low and q is high, the DROC model is less able to capture degree distributions with long upper tails. Moreover, when s is low and q is high, there will be more isolated vertices in a DROC graph since the expected fraction of isolated vertices is at least $(1-s/n)^{n/(q(s-1))}$. In Theorem 19 we show that when s is low and q is high the clustering coefficient is largest. In this regard the DROC model is somewhat limited; it may not be possible to achieve some very high clustering coefficients while simultaneously capturing the upper tail of the degree distribution and avoiding isolated vertices.

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 18. *Let $D \sim \mathcal{D}_\gamma$ be the power law degree distribution defined as follows:*

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

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 17, and therefore can be used to produce a DROC graph.*

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

By varying s and q we can control the clustering coefficient of a $DROC$ graph.

Theorem 19. *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*

$$\mathsf{E}[C(v)] = (1+o(1))\,\frac{\left(\sum_{u\in V} t(u)^2\right)^2}{d^3n^2s}\left((1-e^{-t})^2q^2 + c_tq^3\right),$$

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.

2.3.1 DROC proofs

Proof. (of Theorem 19.) Let v be a vertex with target degree $t = t(v)$, and let k denote the number communities containing v. First we claim $deg(v) \sim Bin\left((s-1)k, \frac{tq}{s}\right)$. Let s be an arbitrary vertex of a community S containing v.

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

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

$$\begin{aligned}
\mathsf{Pr}[N_u] &= \sum_r \frac{\mathsf{Pr}[u \sim v \mid deg(v) = r]\,\mathsf{Pr}[deg(v) = r]}{r} \\
&= \sum_r \frac{\mathsf{Pr}[u \sim v]\,\mathsf{Pr}[deg(v) = r \mid u \sim v]}{r} \\
&= \mathsf{Pr}[u \sim v]\,\mathsf{E}\left[\frac{1}{deg(v)} \mid u \sim v\right] \\
&= (1+o(1))\left(\frac{s}{n}\right)^2\frac{n}{(s-1)q}\frac{t(u)tq}{sd}\left(\frac{1-e^{-tqk}}{tkq}\right) \qquad (9)\\
&= (1+o(1))\,\frac{t(u)\left(1-e^{-tqk}\right)}{qkdn}.
\end{aligned}$$

To see (9), note that by the first claim $\mathsf{E}\left[\frac{1}{deg(v)} \mid u \sim v\right] = \mathsf{E}\left[\frac{1}{X+1}\right]$ where $X \sim Bin\left((s-1)k - 1, \frac{tq}{s}\right)$. Applying Lemma 14 and assuming $s = \omega(1)$, we obtain

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

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

$$\begin{aligned}
\mathsf{E}[C(v) \mid v \text{ in } k \text{ communities}] &= \sum_{u,w} N_u N_w \Pr[u \sim w \mid u \sim v \text{ and } w \sim v] \\
&= \sum_{u,w} \frac{t(u)t(w)\left(1 - e^{-tqk}\right)^2}{(qkdn)^2} \left(\frac{1}{k} + \left(\frac{s}{n}\right)^2 \frac{n}{(s-1)q}\right) \frac{t(u)t(w)q}{sd} \\
&= (1 + o(1))\,\frac{\left(1 - e^{-tqk}\right)^2 \left(\sum_{u \in V} t(u)^2\right)^2}{qd^3k^3n^2s}. \qquad (10)
\end{aligned}$$

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 $f(k) = \frac{(1-e^{-kqt})^2}{k^3}$. 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 $z \in [1/q, k]$

$$\begin{aligned}
f''(z) &= \frac{12(1 - e^{-kqt})^2}{k^5} - \frac{12e^{-kqt}(1 - e^{-kqt})qt}{k^4} + \frac{2e^{-2kqt}q^2t^2}{k^3} - \frac{2e^{-kqt}(1 - e^{-kqt})q^2t^2}{k^3} \\
&\le \frac{12(1 - e^{-kqt})^2}{k^5} + \frac{2e^{-2kqt}q^2t^2}{k^3} \\
&\le q^5\left(12 + 2t^2e^{-2t}\right),
\end{aligned}$$

and

$$f''(z) \ge 0.$$

It follows that

$$f\left(\frac{1}{q}\right) + f'\left(\frac{1}{q}\right)\left(k - \frac{1}{q}\right) \le f(k) \le f\left(\frac{1}{q}\right) + f'\left(\frac{1}{q}\right)\left(k - \frac{1}{q}\right) + q^5\left(6 + t^2e^{-2t}\right)\left(k - \frac{1}{q}\right)^2. \quad (11)$$

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 $f(M)$,

$$\begin{aligned}
\mathsf{E}[f(M)] &\le \mathsf{E}\left[f\left(\frac{1}{q}\right) + f'\left(\frac{1}{q}\right)\left(M - \frac{1}{q}\right) + q^5\left(12 + 2t^2e^{-2t}\right)\left(M - \frac{1}{q}\right)^2\right] \\
&= (1 - e^{-t})^2q^3 + \frac{1}{q}\left(1 - \frac{s}{n}\right)q^5\left(6 + t^2e^{-2t}\right) \\
&\le (1 - e^{-t})^2q^3 + q^4\left(6 + t^2e^{-2t}\right)
\end{aligned}$$

and

$$\mathsf{E}[f(M)] \geq \mathsf{E}\left[f\left(\frac{1}{q}\right) + f'\left(\frac{1}{q}\right)\left(M - \frac{1}{q}\right)\right] = (1-e^{-t})^2 q^3.$$

Therefore $\mathsf{E}[f(M)] = (1-e^{-t})^2 q^3 + c_t q^4$ for some constant $c_t \in [0, 6.2)$.

Finally, we compute

$$\begin{aligned}
\mathsf{E}[C(v)] &= \sum_k \mathsf{Pr}[M=k] \frac{\left(1-e^{-tqk}\right)^2 \left(\sum_{u\in V} t(u)^2\right)^2}{qd^3k^3n^2s} \\
&= \frac{\left(\sum_{u\in V} t(u)^2\right)^2}{qd^3n^2s} \mathsf{E}[f(M)] \\
&= (1+o(1)) \frac{\left(\sum_{u\in V} t(u)^2\right)^2}{d^3n^2s} \left((1-e^{-t})^2 q^2 + c_t q^3\right).
\end{aligned}$$

□

Proof. (of Corollary 18.) Let $d = mean(D)$. We compute

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

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 $t(v_i) > k_0$.

$$\begin{aligned}
\mathsf{Pr}\left[\max_i t(v_i) > t_0\right] &\leq \mathsf{E}[X] = n\,\mathsf{Pr}[t(v_1) > t_0] \leq n \sum_{i=t_0+1}^{\infty} \frac{i^{-\gamma}}{\zeta(\gamma)} \\
&\leq n \int_{i=t_0}^{\infty} \frac{i^{-\gamma}}{\zeta(\gamma)} = \left(\frac{1}{\zeta(\gamma)(\gamma-1)}\right) n t_0^{1-\gamma} = o(1).
\end{aligned}$$

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

2.4 Comparison to other random graph models.

The ROC model captures any pair of triangle-to-edge and four-cycle-to edge ratios simultaneously, and the DROC model can exhibit a wide range of degree distributions with high clustering coefficient. Previous work [18], [29], and [31] 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 range of clustering coefficient. The algorithm presented in [41] produces graphs with tunable degree distribution and clustering, but unlike ROC graphs, there is no underlying community structure and the resultant graphs do not exhibit the commonly observed inverse relationship between degree and clustering coefficient.

The Block Two-Level Erdős and Rényi (BTER) model produces graphs with scale-free degree distributions and random dense communities [33]. However, the communities in the BTER model do not overlap; all vertices are in precisely one E-R community and all other edges are added during a subsequent configuration model phase of construction. Moreover, in the BTER model community membership is determined by degree, which ensures that all vertices in a BTER community have similar degree. In contrast, the degree distribution within a DROC community is a random sample of the entire degree distribution.

Mixed membership stochastic block models have traditionally been applied in settings with overlapping communities [3], [19], [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.

3 Convergent sequences of sparse graphs

In this section, we discuss a few example sequences to illustrate the notions of convergence. Section 3.2 focuses on the convergence of sequences of random graphs. Not all sequences of graphs converge or contain a convergent subsequence according to our defintion; see Section 6.1 for an example.

3.1 Hypercube and rook graphs

We begin with the hypercube sequence, which directly motivates this paper.

Lemma 20 (hypercube limit)**.** *The d-dimensional hypercube is a graph on 2^d vertices, each labeled with a string in $\{0,1\}^d$. Two vertices are adjacent if the Hamming distance of their labels is 1. Let (G_d) be the sequence of d-dimensional hypercubes. The sparsity exponent of the sequence (G_d) is $1/2$ and the sequence is fully convergent with limit $(w_3, w_4, \dots)$ where*

$$w_k = \begin{cases} (k-1)!! & \text{for } k \text{ even} \\ 0 & \text{for } k \text{ odd.} \end{cases}$$

The k-limit of the sequence is $(w_3, w_4, \dots w_k)$.

Proof. We claim that for k even $W_k(G_d) = (k-1)!!nd^{k/2} + o\left(nd^{k/2}\right)$ where $n = 2^d$. Each hypercube edge (u, v) corresponds to a one coordinate difference between the labels of u and v. We think of k-walks on the hypercube as length k strings where the i^{th} character indicates which of the d coordinates is changed on the i^{th} edge of the walk. In closed walks each coordinate that is changed is changed back, so every coordinate appearing in the corresponding string appears an even number of times. Therefore at most $k/2$ coordinates appear in the string. Let Y_i be the number of length k strings with i distinct characters that correspond to a closed k-walk. Since there are d possible coordinates, there are $\binom{d}{i}$ ways to select the i characters and so $Y_i = \Theta(d^i)$. Therefore

$$W_k(G_d) = nY_{k/2} + o\left(nd^{k/2}\right).$$

There are $\binom{d}{k/2}$ ways to select the coordinates to change and $k!/2^{k/2}$ length k strings where $k/2$ characters appear twice. Thus

$$Y_{k/2} = d^{k/2}\frac{k!}{2^{k/2}\left(\frac{k}{2}\right)!} + o\left(d^{k/2}\right) = (k-1)!!d^{k/2} + o\left(d^{k/2}\right),$$

and the claim follows.

Note that there are no odd closed walks in the hypercube because it is bipartite. Therefore $W_k(G) = 0$ for k odd. It follows that the sparsity exponent is $1/2$ and the limit vector is as stated. $\square$

Our second example is a strongly regular family with a different sparsity exponent.

Lemma 21 (rook's graph limit)**.** *The rook graph G_k on k^2 vertices is the Cartesian product of two cliques of size k. (Viewing the vertices as the squares of a $k \times k$ chessboard, the edges represent all legal moves of the rook.) Let (G_k) be the sequence on rook graphs. The sparsity exponent of (G_k) is 1 and the sequence is fully convergent with limit $(w_3, w_4, \dots)$ where $w_j = 2^{2-j}$.*

Proof. The rook's graph is the strongly regular graph on $n = k^2$ vertices with degree $d = 2k-2$ such that each pair of adjacent vertices have $\lambda = k-2$ common neighbors and each pair of non-adjacent vertices have $\mu = 2$ common neighbors. The classical result [10] states that the eigenspectrum of a strongly regular graph is

d with multiplicity 1,

$\frac{1}{2}\left((\lambda-\mu) + \sqrt{(\lambda-\mu)^2 + 4(d-\mu)}\right)$ with multiplicity $\frac{1}{2}\left((n-1) - \frac{2d + (v-1)(\lambda-\mu)}{\sqrt{(\lambda-\mu)^2 + 4(d-\mu)}}\right)$, and

$\frac{1}{2}\left((\lambda-\mu) - \sqrt{(\lambda-\mu)^2 + 4(d-\mu)}\right)$ with multiplicity $\frac{1}{2}\left((n-1) + \frac{2d + (v-1)(\lambda-\mu)}{\sqrt{(\lambda-\mu)^2 + 4(d-\mu)}}\right)$.

Therefore the eigenspectrum of the rook graph G_k is

$2k-2$ with multiplicity 1, -2 with multiplicity $(k-1)^2$, and $k-2$ with multiplicity $2k-2$.

We compute

$$W_j(G_k) = (2k-2)^j + (k-1)^2(-2)^j + (2k-2)(k-2)^j = 2k^{j+1} + O\left(k^j\right).$$

Therefore

$$\lim_{k\to\infty} W_j(G_k, 1) = \lim_{k\to\infty} \frac{2k^{j+1} + O\left(k^j\right)}{k^2(2k-2)^{j-1}} = 2^{2-j}.$$

$\square$

3.2 Almost sure convergence for sequences of random graphs

By an abuse of notation, we say that a sequence of random graphs converges to a limit vector L if a sequence of graphs drawn from the sequence of random graph models almost surely converges L. Lemma 22 gives a method for showing that a sequence of random graphs converges, which we apply to describe the limits of sequence of E-R graphs (Lemma 24). We will again apply Lemma 22 when we discuss the convergence of sequences of ROC graphs (Theorem 34 and Corollary 35).

Definition 7 (convergence of random graph sequences)**.** *Let $M = (M_i)$ be a sequence of random graph models. Let S be a sequence of graphs (S_i) where $S_i \sim M_i$. We say the sequence of random graphs M converges to L if a sequence S drawn from M almost surely converges to L.*

Lemma 22. *Let $M = (M_i)$ be a sequence of random graph models, and let (S_i) be a sequence of graphs where $S_i \sim M_i$. Let $\varepsilon > 0$ and $A_{i,\varepsilon,\alpha}(w_j)$ be the event that $|W_j(S_i, \alpha) - w_j| \geq \varepsilon$.*

1. *If for all j and $\varepsilon > 0$*
$$\sum_{i=1}^{\infty} \mathsf{Pr}[A_{i,\varepsilon,\alpha}(w_j)] < \infty,$$
then M converges to $L = (w_3, w_4, \dots)$ with sparsity exponent α.

2. *If the above hypothesis holds for all $j \leq k$, then M k-converges to $L = (w_3, w_4, \dots w_k)$ with k-sparsity exponent α.*

3. *Let $D(S_i)$ be the random variable for the average degree of a vertex in S_i, let $d_i = \mathsf{E}[D(S_i)]$, and let n_i be the number of vertices of S_i. If $\lim_{i\to\infty} \frac{\mathsf{E}[W_j(S_i)]}{n_i d_i^{1+\alpha(j-2)}} = w_j$ then there exists an index i_0 and a constant C such that*
$$\sum_{i=1}^{\infty} \mathsf{Pr}[A_{i,\varepsilon,\alpha}(w_j)] \leq C + \sum_{i=i_0}^{\infty} \frac{\mathsf{Var}[D(S_i)]}{d_i^2} + \frac{\mathsf{Var}[W_j(S_i)]}{\left(n_i d_i^{1+\alpha(j-2)}\right)^2}.$$

Proof. We begin with (1) and (2). Fix j. To show that $W_j(S_n, \alpha) \to w_j$ almost surely, it suffices to show that for all $\varepsilon > 0$, $\mathsf{Pr}[A_{i,\varepsilon,\alpha}(w_j)$ occurs infinitely often$] = 0$. By the Borel Cantelli Lemma $\sum_{n=1}^{\infty} \mathsf{Pr}[A_{i,\varepsilon,\alpha}(w_j)] < \infty$ implies $\mathsf{Pr}[A_{i,\varepsilon,\alpha}(w_j)$ occurs infinitely often$] = 0$. Statements (1) and (2) follow from the fact that a countable intersection of almost sure events occurs almost surely.

For (3), we apply Lemma 23 which bounds the probability $W_j(S_i, \alpha)$ deviates from expectation by separately bounding the probabilities that the number of edges and the number of closed j-walks in S_i deviate from expectation. Let $g_i = w_j - \frac{\mathsf{E}[W_j(G_i)]}{n_i d_i^{1+\alpha(j-2)}}$, and so $g_i = o(1)$. Let i_0 be such that for all $i \geq i_0$, $|g_i| < \varepsilon/4$, and let $c = \min\{\delta, \varepsilon/4\}$. By Lemma 22(3) for all $i \geq i_0$
$$\mathsf{Pr}[A_{i,\varepsilon,\alpha}(w_j)] \leq \frac{1}{c^2}\left(\frac{\mathsf{Var}[D(G_i)]}{d_i^2} + \frac{\mathsf{Var}[W_j(G_i)]}{\left(n_i d_i^{1+\alpha(j-2)}\right)^2}\right).$$

The claim follows from the observation that
$$\sum_{i=1}^{\infty} \mathsf{Pr}[A_{i,\varepsilon,\alpha}(w_j)] \leq i_0 + \frac{1}{c^2}\sum_{i=i_0}^{\infty} \frac{\mathsf{Var}[D(S_i)]}{d_i^2} + \frac{\mathsf{Var}[W_j(S_i)]}{\left(n_i d_i^{1+\alpha(j-2)}\right)^2}.$$

$\square$

Lemma 23. *Let S be a random graph on n vertices. Let $\varepsilon > 0$ and $A_{\varepsilon,\alpha}(w_j)$ be the event that $|W_j(S,\alpha) - w_j| \geq \varepsilon$. Let $D(S)$ be the random variable for the average degree of a vertex in S, and let $d = \mathsf{E}[D(S)]$. Let $g = w_j - \frac{\mathsf{E}[W_j(S)]}{nd^{1+\alpha(j-2)}}$. For $|g| < \varepsilon/2$, $\delta = \min\left\{\frac{\varepsilon}{2j(w_j+\varepsilon)}, \frac{1}{2(j-1)^2}\right\}$ and $\lambda = \varepsilon/2 - |g|$,*

$$\mathsf{Pr}[A_{\varepsilon,\alpha}(w_j)] \leq \frac{\mathsf{Var}[D(S)]}{\delta^2 d^2} + \frac{\mathsf{Var}[W_j(S)]}{\lambda^2 \left(nd^{1+\alpha(j-2)}\right)^2}.$$

Proof. Observe that if $A_{i,\varepsilon,\alpha}(w_j)$ holds, then for any $\delta > 0$ at least one of the following events hold:

(a) $|D(S) - d| > \delta d$

(b) $W_j(S) \geq (w_j + \varepsilon)\left(n\left(d(1-\delta)\right)^{1+\alpha(j-2)}\right)$

(c) $W_j(S) \leq (w_j - \varepsilon)\left(n\left(d(1+\delta)\right)^{1+\alpha(j-2)}\right)$.

When (a) does not hold

$$\frac{W_j(S)}{n\left(d(1+\delta)\right)^{1+\alpha(j-2)}} \leq W_j(S,\alpha) \leq \frac{W_j(S)}{n\left(d(1-\delta)\right)^{1+\alpha(j-2)}}.$$

Assume (a) does not hold and $A_{\varepsilon,\alpha}(w_j)$. If $W_j(S,\alpha) \geq w_j + \varepsilon$ then (b) holds. If $W_j(S,\alpha) \leq w_j - \varepsilon$ then (c) holds. The observation follows.

We now give a bound on the probability of (b) or (c). Let $\gamma^- = 1 - (1-\delta)^{1+\alpha(j-2)}$ and $\gamma^+ = (1+\delta)^{1+\alpha(j-2)} - 1$. We write $w_j = \frac{\mathsf{E}[W_j(S)]}{nd^{1+\alpha(j-2)}} + g$. Statement (b) becomes

$$W_j(S) - \mathsf{E}[W_j(S)] \geq \left(\varepsilon + g - \gamma^-(w_j + \varepsilon)\right) nd^{1+\alpha(j-2)},$$

and statement (c) becomes

$$W_j(S) - \mathsf{E}[W_j(S)] \leq \left(\gamma^+(w_j - \varepsilon) - \varepsilon + g\right) nd^{1+\alpha(j-2)}.$$

Under the assumptions that $\delta = \min\left\{\frac{\varepsilon}{2j(w_j+\varepsilon)}, \frac{1}{2(j-1)^2}\right\}$ and $\alpha \leq 1$,

$$\gamma^- = 1 - (1-\delta)^{1+\alpha(j-2)} \leq \delta(1 + \alpha(j-2)) < \delta j \leq \frac{\varepsilon}{2(w_j+\varepsilon)}$$

$$\gamma^+ = (1+\delta)^{1+\alpha(j-2)} - 1 \leq \delta(j-1) + \sum_{i=2}^{j-1} \delta^i \binom{j-1}{i} \leq \delta(j-1) + 2\delta^2(j-1)^2 < \delta j \leq \frac{\varepsilon}{2(w_j - \varepsilon)}.$$

Let $\lambda = \varepsilon/2 - |g|$ and note

$$\varepsilon + g - \gamma^-(w_j+\varepsilon) \geq \lambda \quad \text{and} \quad \varepsilon - g - \gamma^+(w_j - \varepsilon) \geq \lambda.$$

It follows from Chebyshev's inequality that

$$\mathsf{Pr}[(b) \text{ or } (c)] \leq \mathsf{Pr}\left[|W_j(S) - \mathsf{E}[W_j(S)]| \geq c(\delta) nd^{1+\alpha(j-2)}\right] \leq \frac{\mathsf{Var}[W_j(S)]}{\left(\lambda nd^{1+\alpha(j-2)}\right)^2}.$$

Finally, we apply Chebyshev's inequality to bound the probability of (a), apply a union bound for the event $A_{i,\varepsilon,\alpha}(w_j)$, and obtain

$$\Pr[A_{\varepsilon,\alpha}(w_j)] \leq \Pr[(a)] + \Pr[(b) \text{ or } (c)] \leq \frac{\mathrm{Var}[D(S)]}{\delta^2 d^2} + \frac{\mathrm{Var}[W_j(S)]}{\lambda^2 \left(nd^{1+\alpha(j-2)}\right)^2}.$$

□

Our final example is sequences of Erdős-Rényi random graphs. These demonstrate some of the subtler issues with defining limits.

Lemma 24 (Erdős-Rényi sequence)**.** *Let $(G_n) \sim G\left(n^{2\ell}, n^{2-2\ell}\right)$ for $\ell > 1$. We denote the j^{th} Catalan number $Cat_j = \frac{1}{j+1}\binom{2j}{j}$.*

1. *For $k < 2\ell$, the k-sparsity exponent of (G_n) is $1/2$ and the k-limit is $(w_3, w_4, \dots w_k)$ where $w_j = 0$ for i odd and $w_j = Cat_{i/2}$ for i even.*
2. *For $k = 2\ell$, the k-sparsity exponent of (G_n) is $1/2$ and the k-limit is $(w_3, w_4, \dots, w_{k-1}, \overline{w}_k)$ where $w_j = 0$ for odd j, $w_j = Cat_{j/2}$ for even j, and $\overline{w}_k = w_k + 1$.*
3. *For $k > 2\ell$, the sparsity exponent of (G_n) is $\frac{k-\ell-1}{k-2}$ and the k-limit is $(w_3, w_4, \dots, w_k)$ where $w_j = 0$ for $j < k$, $w_k = 1$.*
4. *The sparsity exponent of (G_n) is 1 and the limit is $(0, 0, \dots)$.*

First we compute the expectation and variance of the number of closed i-walks in a E-R random graph.

Lemma 25. *Let $G \sim G(n, d/n)$. Let $W_j(G)$ be the random variable for the number of closed j walks in G. Then*

$$\mathrm{E}[W_j(G)] = d^j + Cat_{j/2} n d^{\lfloor j/2 \rfloor} + \Theta\left(d^{j-1} + nd^{\lfloor j/2 \rfloor - 1}\right)$$

$$\mathrm{Var}[W_j(G)] = \Theta\left(d^{2j-1} + n^2 d^{2\lfloor j/2 \rfloor - 1} + nd^{\lfloor j/2 \rfloor + j - 1}\right).$$

Proof. Let $W_j^{a,b}(G_n)$ be the number of closed j walks involving a vertices and b edges, so $b \leq j$ and either $a \leq b$ (the walk contains a cycle) or $a = b + 1$ and $b \leq j/2$ (the walk traces a tree). Let $f(a, b, j)$ be the number of closed j-walks with b total edges on a labeled vertices $1, 2, \dots a$ such that the order in which the vertices are first visited is $1, 2, \dots a$. Note $f(j, j, j) = 1$ and $f(j/2 + 1, j/2, j) = Cat_{j/2}$ for j even (because there are Cat_b ordered trees on b edges, see [36]). Let $\zeta(j)$ be one if j is even and zero otherwise. We split the sum based on whether the walk contains a cycle or traces a tree and compute

$$\begin{aligned}
\mathrm{E}[W_j(G)] &= \sum_{b=1}^{j} \sum_{a=1}^{b+1} \mathrm{E}\left[W_j^{a,b}(G)\right] = \sum_{b=1}^{j} \sum_{a=1}^{b+1} f(a, b, j) \frac{n!}{(n-a)!} \left(\frac{d}{n}\right)^b \\
&= \sum_{b=3}^{j} \sum_{a=1}^{b} f(a, b, j) \frac{n!}{(n-a)!} \left(\frac{d}{n}\right)^b + \sum_{b=1}^{\lfloor j/2 \rfloor} f(b+1, b, j) \frac{n!}{(n-(b+1))!} \left(\frac{d}{n}\right)^b \\
&= d^j + \zeta(j) Cat_{j/2} n d^{j/2} + \Theta\left(d^{j-1} + nd^{\lfloor j/2 \rfloor - 1}\right).
\end{aligned}$$

To find the variance of $W_j(G)$ we compute the expectation squared. Let $P_j^{a,b}(G)$ be the number of pairs of closed j walks involving a total of a vertices and b edges. Let $g(a,b,j)$ be the number of pairs of closed j-walks with b total edges on a labeled vertices $1,2,\dots a$ such that the order in which the vertices are first visited is $1,2,\dots a$ when the first walk is traversed then the second walk. Note $g(2j,2j,j)=1$ and $g(j+2,j,j)=\left(Cat_{j/2}\right)^2$ (since there are $(Cat_b)^2$ ways to pick two disjoint ordered trees on b edges).

We split the sum based on whether both walks contain a cycle, or both trace trees, or one traces a tree and one traces a cycle and compute

$$\begin{aligned}
\mathsf{E}\left[W_j(G)^2\right] &= \sum_{b=1}^{2j}\sum_{a=1}^{b+2}\mathsf{E}\left[P_j^{a,b}(G)\right] = \sum_{b=1}^{2j}\sum_{a=1}^{b+2} g(a,b,j)\frac{n!}{(n-a)!}\left(\frac{d}{n}\right)^b \\
&= \sum_{b=1}^{2j}\sum_{a=1}^{b} g(a,b,j)\frac{n!}{(n-a)!}\left(\frac{d}{n}\right)^b + \sum_{b=1}^{2\lfloor j/2\rfloor} g(b+2,b,j)\frac{n!}{(n-(b+2))!}\left(\frac{d}{n}\right)^b \\
&\quad + \sum_{b=1}^{2j}\sum_{a=1}^{b+1} g(a,b,j)\frac{n!}{(n-a)!}\left(\frac{d}{n}\right)^b \\
&= d^{2j} + \zeta(j)\left(Cat_{j/2}\right)^2 n^2 d^j + \Theta\left(d^{2j-1} + n^2 d^{2\lfloor j/2\rfloor -1} + nd^{\lfloor j/2\rfloor + j-1}\right).
\end{aligned}$$

It follows

$$\mathsf{Var}[W_j(G)] = \mathsf{E}\left[W_j(G)^2\right] - \mathsf{E}[W_j(G)]^2 = \Theta\left(d^{2j-1} + n^2 d^{2\lfloor j/2\rfloor -1} + nd^{\lfloor j/2\rfloor + j-1}\right).$$

□

Now we use these computations and apply Lemma 22 to prove Lemma 24.

Proof. (of Lemma 24). By Lemma 25

$$\mathsf{E}[W_j(G_n)] = \begin{cases} n^{2j} + Cat_{j/2} n^{2\ell+j} + o\left(n^{2j} + n^{2\ell+2\lfloor j/2\rfloor}\right) & j \text{ is even} \\ n^{2j} + o\left(n^{2j} + n^{2\ell+2\lfloor j/2\rfloor}\right) & j \text{ is odd.}\end{cases}$$

We compute the k-sparsity exponent

$$\alpha_k = \inf_{a\in[1/2,1]}\left\{a \mid \mathsf{E}[W_j(G_n)] = O\left(n^{2\ell+2+2\alpha(j-2)}\right) \text{ for all } j\le k\right\} = \max\left\{\frac{1}{2}, \frac{k-\ell-1}{k-2}\right\},$$

and the sparsity exponent

$$\alpha = \inf_{a\in[1/2,1]}\left\{a \mid \mathsf{E}[W_j(G_n)] = O\left(n^{2\ell+2+2\alpha(j-2)}\right) \text{ for all } j\right\} = 1.$$

Note for each of the cases outlined in the statement, $w_j = \lim_{i\to\infty}\frac{\mathsf{E}[W_j(G_n)]}{n^{2\ell+2+2\alpha(j-2)}}$ where α is the corresponding sparsity exponent. To prove convergence for cases 1-3, we apply Lemma 22(2) and for case 4 we apply Lemma 22(1). By Lemma 22(3), it remains to show that

$$\sum_{n=n_0}^{\infty}\left(\frac{\mathsf{Var}[D(G_n)]}{d_n^2} + \frac{\mathsf{Var}[W_j(G_n)]}{\left(n^{2\ell+2+2\alpha(j-2)}\right)^2}\right) < \infty. \tag{12}$$

Note $D(G_n) \sim Bin\left(\binom{n^{2\ell}}{2}, n^{2-2\ell}\right)$, and so $\mathsf{Var}[D(G_n)] = \binom{n^{2\ell}}{2} n^{2-2\ell}\left(1 - n^{2-2\ell}\right)$ and $d_n = \binom{n^{2\ell}}{2} n^{2-2\ell}$. It follows

$$\sum_{n=2}^{\infty} \frac{\mathsf{Var}[D(G_n)]}{d_n^2} = \sum_{n=2}^{\infty} \frac{\binom{n^{2\ell}}{2} n^{2-2\ell}\left(1 - n^{2-2\ell}\right)}{\left(\binom{n^{2\ell}}{2} n^{2-2\ell}\right)^2} \leq \sum_{n=2}^{\infty} 2n^{-2\ell-2} < \infty.$$

We show the sum of the variance term is finite by considering the cases separately. By Lemma 25,

$$\mathsf{Var}[W_j(G_n)] = \Theta\left(n^{4j-2} + n^{4\ell+4\lfloor j/2\rfloor-2} + n^{2\ell+2\lfloor j/2\rfloor+2j-2}\right),$$

and so

$$X := \frac{\mathsf{Var}[W_j(G_n)]}{\left(n^{2\ell+2+2\alpha(j-2)}\right)^2} = \Theta\left(n^{4j-6-4\ell-4\alpha(j-2)} + n^{-2+(j-2)(2-4\alpha)} + n^{-2\ell+(j-2)(3-4\alpha)}\right).$$

For (1) and (2), $\alpha = 1/2$, $j \leq k \leq 2\ell$, and so $X = \Theta\left(n^{2j-4\ell-2} + n^{-2} + n^{-2\ell+j-2}\right) = O\left(n^{-2}\right)$. For (3), $\alpha = \frac{k-\ell-1}{k-2}$, $k > 2\ell$, and $j \leq k$. Since $\alpha \geq \frac{j-\ell-1}{j-2}$, $\Theta\left(n^{4j-6-4\ell-4\alpha(j-2)}\right) = O\left(n^{-2}\right)$. Since $\alpha > 1/2$, $\Theta\left(n^{-2+(j-2)(2-4\alpha)}\right) = O\left(n^{-2}\right)$. Since $\alpha > 1/2$ and $k > 2\ell$, $\Theta\left(n^{-2\ell+(j-2)(3-4\alpha)}\right) = O\left(n^{-2\ell+k-2}\right) = O\left(n^{-2}\right)$. It follows that $X = O\left(n^{-2}\right)$. For (4) $\alpha = 1$ and $j \geq 3$, and so $X = \Theta\left(n^{2-4\ell} + n^{2-2j} + n^{-2\ell-j+2}\right) = O\left(n^{-2}\right)$. Therefore in all cases

$$\sum_{n=2}^{\infty} \frac{\mathsf{Var}[W_j(G_n)]}{\left(n^{2\ell+2+2\alpha(j-2)}\right)^2} = \sum_{n=2}^{\infty} O\left(n^{-2}\right) < \infty,$$

and the statement follows from (12). □

4 Approximating a convergent sequence by a ROC

4.1 A parameterization of the ROC model

In this section, we introduce a parametrization of the ROC model which will be particularly convenient in proofs. The distribution $\mathcal{D}$ of a ROC family is specified by a number $a \in [0, 1]$ and a distribution μ on triples (m_i, q_i, β_i) with probability μ_i for the i^{th} triple. Communities are generated by repeatedly picking a triple from the distribution μ. When $\beta_i = 0$, the community has expected size $s = m_i d^a$ and density q_i. If $\beta_i = 1$, indicating balanced bipartiteness, the community is defined on a bipartite graph with $m_i d^a$ vertices expected in each class.

ROC(n, d, μ, a).

Input: number of vertices n, degree d, $a \in [0,1]$, and μ a distribution on a finite set of triples (m_i, q_i, β_i) where (m_i, q_i, β_i) is selected with probability μ_i, $m_i > 0$, $\sum_i \mu_i = 1$, $\beta_i \in \{0,1\}$, $0 \le q_i \le 1$, and $\max m_i d^a \le n$. Let B be the set of indices i such that $\beta_i = 1$ and B^c be the set of indices i such that $\beta_i = 0$. Let $x = 1/\left(\sum_{i \in B^c} \mu_i m_i^2 q_i + 2\sum_{i\in B} \mu_i m_i^2 q_i\right)$.

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

Repeat xnd^{1-2a} times:

1. Randomly select a pair (m_i, q_i, β_i) from μ with probability μ_i.
2. If $\beta_i = 0$
   (a) Pick a random subset S of vertices (from $\{1, 2, \ldots, n\}$) by selecting each vertex independently with probability $m_i d^a/n$.
   (b) Add the random graph $G_{|S|,q_i}$ on S, i.e., for each pair in S, add the edge between them independently with probability q_i; if the edge already exists, do nothing.

   If $\beta_i = 1$
   (a) Pick a random subset S of vertices (from $\{1, 2, \ldots, n\}$) by selecting each vertex independently with probability $2m_i d^a/n$. For each vertex that is in S randomly assign it to either S_1 or S_2.
   (b) Add the bipartite random graph $G_{|S_1|,|S_2|,q_i}$ on S, i.e., for each pair $u \in S_1$ and $v \in S_2$, add the edge between them independently with probability q_i; if the edge already exists, do nothing.

The parameters (n, d, μ, a) are valid ROC parameters if the conditions described under *input* in the box hold. A ROC family $\mathcal{D} = (\mu, a)$ refers to the set of ROC models with parameters μ and a and any valid n and d.

The sparsity exponent of a sequence determines the parameter a of the ROC family that achieves the limit vector. If a vector is achievable with sparsity exponent α then the ROC family that achieves the vector will have parameter $a = \alpha$ unless $\alpha = 1/2$ and vector is the Catalan vector ($w_j = 0$ for j odd and $w_j = Cat_{j/2}$ for j even). In this case any ROC family with $a < 1/2$ achieves the vector.

4.2 Cycles, walks, and limits of ROC's

In this section we describe a combinatorial relationship between closed walk counts and simple cycle counts that appears in graphs in which each vertex is in approximately the same number of simple cycles (Definition 8). Throughout this paper, for convenience we refer to a simple k-cycle as a k-cycle. For example, under this convention a k-cycle graph has $2k$ k-cycles because there are $2k$ distinct closed walks that traverse a k-cycle.

Definition 8 (cycle-walk transform)**.** *Let*

$$\mathcal{S}_k = \left\{ \{(a_1, t_1), (a_2, t_2), \ldots, (a_j, t_j)\} \mid \sum_{i=1}^{j} a_i t_i = k, a_i \neq a_j \text{ for } i \neq j, a_i, t_i \in \mathbb{Z}^+, a_i > 1 \right\}.$$

Define $T((c_3, c_4, \ldots c_n)) = (w_3, w_4, \ldots w_n)$ *as the invertible transform*

$$w_k = \sum_{S \in \mathcal{S}_k} \frac{k!}{(\prod t_i!)(k+1-\sum t_i)!} \prod_{i=1}^{j} (c_{a_i})^{t_i}.$$

The transform T is analogously defined for infinite count vectors.

Remark 26. *The first few terms of T are illustrated below:*

$$\begin{aligned}
w_3 &= c_3 \\
w_4 &= 2 + c_4 \\
w_5 &= c_5 + 5c_3 \\
w_6 &= c_6 + 6c_4 + 3c_3^2 + 5 \\
w_7 &= 21c_3 + 7c_3c_4 + c_7 + 7c_5 \\
w_8 &= 8c_3c_5 + 28c_3^2 + c_8 + 8c_6 + 28c_4 + 4c_4^2 + 14 \\
w_9 &= 9c_3c_6 + 84c_3 + 12c_3^3 + 9c_7 + 36c_5 + 9c_4c_5 + c_9 + 72c_3c_4 \\
w_{10} &= 42 + 5c_5^2 + 180c_3^2 + 45c_3^2c_4 + 10c_4c_6 + 90c_3c_5 + 10c_8 + c_{10} + 45c_6 + 10c_3c_7 + 120c_4 + 45c_4^2.
\end{aligned}$$

We see that T is invertible by using induction to show that each c_j is completely determined by the vector $(w_3, \ldots w_j)$. Note $c_3 = w_3$ is completely determined. Assume $c_3, \ldots c_j$ have been completely determined. Note that $w_{j+1} = c_{j+1} + f((c_3, c_4, \ldots c_j))$ for some function f. Since w_{j+1} is given and f is a function of values that are already determined, there is only one choice for c_{j+1}.

In Section 4.2.1 we derive the coefficient of $\prod_{i=1}^{j}(c_{a_i})^{t_i}$ in T by counting the number of walk structures that can be decomposed into $t_1, t_2, \ldots t_j$ cycles of lengths $a_1, a_2, \ldots a_j$ respectively. In Section 4.2.2, we define class of *locally regular* graphs in which each vertex is in the same number of cycles, and then show this class of graphs exhibits the relationship between cycles and closed walks given in Definition 8.

In Section 4.2.3, we prove Theorem 27, which describes the limit achieved by a ROC family (μ, a). The parameter a plays an important role. When $a < 1/2$, the closed walks that trace trees dominate the closed walk count, so the limit is the Catalan sequence. When $a > 1/2$ the closed walks that trace simple cycles dominate the closed walk count, so the limit is the normalized number of expected simple cycles. However, when $a = 1/2$, cycles, trees, and other walk structures are all of the same order, and so the relationship given in Definition 8 appears in the limit.

Theorem 27. *Let (μ, a) be a ROC family. Let B be the set of all i such that $\beta_i = 1$, let B^c be the set of all i such that $\beta_i = 0$, and let $x = 1/\left(\sum_{i \in B^c} \mu_i m_i^2 q_i + 2\sum_{i \in B} \mu_i m_i^2 q_i\right)$. Define*

$$c(k) = \begin{cases} 1 & k = 2 \\ x \sum_{i \in B^c} \mu_i (m_i q_i)^k & k \text{ odd and } k \geq 3 \\ x \sum_{i \in B^c} \mu_i (m_i q_i)^k + 2x \sum_{i \in B} \mu_i (m_i q_i)^k & k \text{ even and } k \geq 4. \end{cases}$$

Let $Cat_n = \frac{1}{n+1}\binom{2n}{n}$ denote the n^{th} Catalan number, and let T be as given in Definition 8.

1. *If $a < 1/2$, the ROC family fully achieves the limit $(0, Cat_2, 0, Cat_3, \dots)$ with sparsity exponent $1/2$.*
2. *If $a = 1/2$, the ROC family fully achieves the limit $T((c(3), c(4), c(5), \dots))$ with sparsity exponent $1/2$.*
3. *If $a > 1/2$, the ROC family fully achieves the limit $(c(3), c(4), c(5), \dots)$ with sparsity exponent a.*

The ROC family achieves the corresponding length $k-2$ prefix as its k-limit with the same k-sparsity exponent for $k \geq 4$.

In Section 4.2.4, we show that the probability the normalized walk count $W_j(G, \alpha)$ of a ROC graph $G \sim ROC(n, d, \mu, a)$ deviates from w_j in the limit achieved by the family vanishes as d grows (Theorem 34). Corollary 35 gives conditions that guarantee that a sequence of graphs drawn from a common ROC family converges almost surely to the limit achieved by the family.

4.2.1 The cycle structure of closed walks

In order to count the number of closed walks in a graph, we divide the closed walks into classes based on the structure of the cycles appearing in the closed walk and then count the number of closed walks in each class. Each class is defined by a "cycle permutation" in which each non-zero character represents the first step of a cycle within the walk and each zero represents a step in a cycle that has already begun (Definition 9). In Lemma 28, we show that the number of cycle permutations corresponding to a walk made of $t_1, t_2, \dots t_j$ cycles of lengths $a_1, a_2, \dots a_j$ respectively is the coefficient of $\prod_{i=1}^{j} (c_{a_i})^{t_i}$ in the cycle-walk transform.

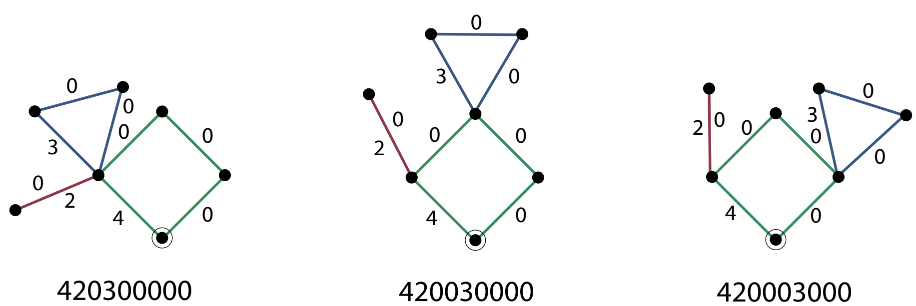

Figure 5: The above walks begin and end at the circled vertex and proceed left to right. Each is labeled with its cycle permutation.

Definition 9 (cycle permutation)**.** *Follow the procedure below to label each step of a closed k-walk $\mathcal{W} = (r_1, r_2, \dots r_k)$ with a label and define the "cycle permutation" P of $\mathcal{W}$ as the labels of the steps in order of traversal.*

1. *Repeat until all steps are labeled:*

Traverse $\mathcal{W}$ skipping a step r_i if it has already been labeled. Let u be the first repeated vertex on this traversal. The modified walk must have traversed a cycle $r_i, r_{i+1}, \dots r_{i+j-1}$ starting at u. Label the first step r_i with the length of the cycle. Label all other steps with zero.

2. *Traverse $\mathcal{W}$ and let P be the string of labels of the steps as they are traversed.*

The following lemma enumerates the cycle permutations using bijection between cycle permutations and generalized Dyck paths (Definition 10).

Lemma 28. *Let $\mathcal{S}_k$ be as given in Definition 8. For each $S \in \mathcal{S}_k$, let M_s be the multiset where a_i appears t_i times and there are $k - \sum_i t_i$ zeros. Let P_S be the set of all permutations of M_s such that the following property holds for all $2 \leq i \leq k+1$:*

$$\sum_{\ell \in N(s)} (\ell - 1) \geq z_i$$

where $N(s)$ is the multiset of non-zero labels that appear before the i^{th} label of the permutation and z_i is the number of times zero occurs before the i^{th} label of the permutation. The set of cycle permutations is $\bigcup_{s \in S} P_S$ and

$$|P_S| = \frac{k!}{(\prod t_i!)(k+1-\sum t_i)!}.$$

To compute the size of P_S in the above lemma we use a bijection between permutations in P_S and generalized sub-diagonal Dyck paths, whose cardinality is given in Lemma 29.

Definition 10 (generalized Dyck path, see [32])**.** *A generalized Dyck path p is a sequence of n vertical steps of height one and $k \leq n$ horizontal steps with positive integer lengths $\ell_1, \ell_2, \dots \ell_k$ satisfying $\sum_{i=1}^{k} \ell_i = n$ on a $n \times n$ grid such that no vertical step is above the diagonal.*

Lemma 29 (from [32])**.** *Let D the set of generalized Dyck paths on a $(k - \sum t_i) \times (k - \sum t_i)$ grid that are made up of t_i horizontal steps of length $a_i - 1$ and $k - \sum t_i$ vertical steps of length 1. Then*

$$|D| = \frac{k!}{(\prod t_i!)(k+1-\sum t_i)!}.$$

Proof. (of Lemma 28) First we show the set of cycle permutations is $\bigcup_{s \in S} P_S$. Let P be a cycle permutation of some closed walk $\mathcal{W}$ of length k. Let $s = \{(a_1, t_1), (a_2, t_2), \dots (a_j, t_j)\} \in S$ where $\{a_1, a_2, \dots a_j\}$ are the non-zero labels of P and each a_i appears t_i times in P (so $\sum a_i t_i = k$). To see that P is in P_S we show that for all $2 \leq i \leq k+1$:

$$\sum_{\ell \in N(s)} (\ell - 1) \geq z_i$$

where $N(s)$ is the multiset of non-zero labels that appear before the i^{th} label of P and z_i is the number of times zero occurs before the i^{th} label of P. Before the i^{th} step of the walk suppose non-zero labels $\ell_1, \dots \ell_k$ have been traversed. The only steps labeled with a zero that have been traversed must be part of a cycle corresponding to one of the labels $\ell_1, \dots \ell_k$. Since ℓ_i labels a cycle of length ℓ_i at most $\sum(\ell_i - 1)$ zero steps have been traversed. Thus $P \in P_S$.

Next we claim that any $P \in \bigcup P_S$ corresponds to a closed walk $\mathcal{W}$. Let $T(s) = \sum t_i$ be the number of non-zero values in each permutation in P_S. We show that for any $k = \sum a_i t_i$ all permutations in P_S correspond to closed walks by induction on $T(s)$. Note for any k and $T(s) = t_1 = 1$, there is one permutation in P_S, $k = a_1$ followed by $k-1$ zeros. This permutation corresponds to a k-cycle. Assume that if $T(s') < T(s)$ then each string in $P_{s'}$ corresponds to a closed walk. We show each $P \in P_S$ corresponds to a closed walk. Consider the last non-zero value of the permutation P. Without loss of generality, suppose this value is a_j and that it occurs at the i^{th} coordinate of P. Since $P \in P_S$, there must be at least $a_j - 1$ zeros to the right of a_j. Removing a_j and $a_j - 1$ zeros to its right in P yields a valid sequence $P' \in P_{s'}$ where $s' = (a_1, \ldots a_j, t_1, \ldots t_{j-1}, t_j - 1)$ and $k' = k - a_j$. By the inductive hypothesis, P' corresponds to a closed walk $\mathcal{W}'$ of length $\sum a_i t_i - a_j$. Add a cycle of length a_j in between the $(i-1)^{st}$ and i^{th} steps of $\mathcal{W}'$ to obtain a closed walk $\mathcal{W}$ of length $k = \sum a_i t_i$. We have shown that $\bigcup P_S$ is the set of all cycle permutations for closed walks.

We compute the size of $|P_S|$ by constructing a bijection between permutations in P_S and a set of subdiagonal generalized Dyck paths. Let $s = \{(a_1, t_1), (a_2, t_2), \ldots (a_j, t_j)\} \in S$. Let D the set of subdiagonal generalized Dyck paths on a $(k - \sum t_i) \times (k - \sum t_i)$ grid that are made up of t_i horizontal steps of length $a_i - 1$ and $k - \sum t_i$ vertical steps of length 1. Each non-zero label a_i of P_S corresponds to a horizontal step of length $a_i - 1$ and each zero label of P_S corresponds to a vertical step of length one. Consider the map between permutations and generalized Dyck paths based on this correspondence. The condition that for all $2 \le i \le k+1$ $\sum_{\ell \in N(s)} (\ell - 1) \ge z_i$ translates to the generalized Dyck path not crossing the diagonal. Thus, the correspondence is a bijection between P_S and D. Lemma 29 implies that $|P_S| = |D| = \frac{k!}{(\prod t_i!)(k+1-\sum t_i)!}$. □

4.2.2 Walk and cycle counts in locally regular graphs

We show that the polynomial relating cycles and closed walks given in Definition 8 governs the relationship between cycles and closed walks in graphs where each vertex is in approximately the same number of cycles.

Definition 11 (k-locally regular, essentially k-locally regular)**.** *Let $C_k(G, v)$ denote the the number of k-cycles at vertex v in G.*

1. *A graph G is k-locally regular if it is regular and $C_j(G, v) = C_j(G, u)$ for all $u, v \in V(G)$ and $j \le k$.*
2. *A sequence of graphs (G_i) with $d_i \to \infty$ and k-sparsity exponent a is essentially k-locally regular if $C_j(G_i, v) - C_j(G_i, u) = o\left(d_i^{j/2}\right)$ for all $u, v \in V(G)$ and $j \le k$.*

Theorem 30. *Let G be a k-locally regular graph on n vertices with degree d. Let $c_k = C_k(G)/(nd^{k/2})$ and $w_k = W_k(G)/(nd^{k/2}) = W_k(G, 1/2)$ where $W_k(G)$ and $C_k(G)$ denote the number of closed k-walks and simple k-cycles in G respectively. Then*

$$w_k = T((c_3, c_4, \ldots c_k)).$$

Proof. We count the number closed walks in G at a vertex v by partitioning the closed walks into sets based on their cycle permutation and computing the size of each partition class. Let $S = \{(a_1, t_1), (a_2, t_2), \ldots, (a_j, t_j)\} \in \mathcal{S}_k$ and $P \in P_S$ as defined in Lemma 28. Define X_P as the number of walks with cycle permutation P at v in G. Let $t = \sum t_i$ be the number of non-zero values in P and let $N(P,i)$ denote the i^{th} non-zero value of the string P. Let $C_k(G,u)$ denote the number of k-cycles at u. Since G is locally regular $C_k(G,u) = c_k d^{k/2}$ for all k. It follows

$$X_P = \prod_{\ell=1}^{t} C_{N(P,\ell)}(G,u) = \prod_{i=1}^{j} \left(c_{a_i} d^{a_i/2}\right)^{t_i} = d^{k/2} \prod_{i=1}^{j} (c_{a_i})^{t_i}. \tag{13}$$

Summing over all $P \in P_S$ and all vertices we obtain

$$W_k(G) = nd^{k/2} \sum_{S \in \mathcal{S}} |P_S| \prod_{i=1}^{j} (c_{a_i})^{t_i}, \quad \text{equivalently} \quad w_k = \sum_{S \in \mathcal{S}} |P_S| \prod_{i=1}^{j} (c_{a_i})^{t_i}.$$

The statement follows directly from Lemma 28. □

Theorem 31. *Let (G_r) be a sequence of essentially k-locally regular graphs with n_r vertices and degree $d_r \to \infty$, sparsity exponent $1/2$, and k-limit $(w_3, w_4, \dots w_k)$. Let $C_j(G_r)$ be the number of j-cycles in G_r. Then $(w_3, w_4, \dots w_k) = T((c_3, c_4, \dots c_k))$ where*

$$c_j = \lim_{s \to \infty} \frac{C_j(G_r)}{n_r d_r^{j/2}}.$$

Proof. We follow the proof of Theorem 30 until line (13). Since the G_s is approximately locally regular rather than locally regular, we have the weaker guarantee that $C_k(G_i,u) = \frac{C_k(G_i)}{n_i} + o\left(d_i^{j/2}\right)$. It follows that for $G = G_r$, $n = n_r$, and $d = d_r$,

$$X_P = \prod_{\ell=1}^{t} C_{N(P,\ell)}(G,u) = \prod_{i=1}^{j} \left(\frac{C_{a_i}(G)}{n} + o\left(d^{a_i/2}\right)\right)^{t_i} = d^{k/2} \prod_{i=1}^{j} \left(\frac{C_{a_i}(G)}{nd^{a_i/2}}\right)^{t_i} + o\left(d^{k/2}\right).$$

Summing over all $P \in P_S$ and all vertices we obtain

$$W_k(G_r) = n_r d_r^{k/2} \sum_{S \in \mathcal{S}} |P_S| \prod_{i=1}^{j} \left(\frac{C_{a_i}(G_r)}{n_r d_r^{a_i/2}}\right)^{t_i} + o\left(n_r d_r^{k/2}\right).$$

Therefore

$$w_k = \lim_{r \to \infty} \frac{W_k(G_r)}{n_r d_r^{k/2}} = \lim_{r \to \infty} \sum_{S \in \mathcal{S}} |P_S| \prod_{i=1}^{j} \left(\frac{C_{a_i}(G_r)}{n_r d_r^{a_i/2}}\right)^{t_i} + o_r(1) = \sum_{S \in \mathcal{S}} |P_S| \prod_{i=1}^{j} (c_{a_i})^{t_i},$$

and the statement follows directly from Lemma 28. □

4.2.3 Limits achieved by ROC families

We prove Theorem 27, which describes the limits of ROC families. The following lemma gives the expected number of closed walks by permutation type.

Lemma 32. *Let $S = \{(a_1, t_1), (a_2, t_2), \ldots, (a_j, t_j)\} \in \mathcal{S}_k$ as defined in Lemma 28, and let $t = \sum_{i=1}^{j} t_i$. Let $X_S(G)$ be the random variable for the number of walks with a permutation type in P_S in $G \sim ROC(n, d, \mu, a)$ where $d = o\left(n^{1/((1-a)k+2a-1)}\right)$. Then for the function c as given in Theorem 27*

1. *For $a < 1$, $\mathsf{E}[X_S(G)] = |P_S| \left(\prod_{i=1}^{j} c(a_i)^{t_i}\right) nd^{(1-2a)t+ak} + o\left(nd^{(1-2a)t+ak}\right)$.*
2. *For $a = 1$ and $t = 1$, $\mathsf{E}[X_S(G)] = |P_S| \left(\prod_{i=1}^{j} c(a_i)^{t_i}\right) nd^{(1-2a)t+ak} + o\left(nd^{(1-2a)t+ak}\right)$.*
3. *For $a = 1$ and $t > 1$, $\mathsf{E}[X_S(G)] = \Theta\left(nd^{(1-2a)t+ak}\right)$.*

Taking $S = \{(k, 1)\}$ in the above lemma gives the number of simple k-cycles in a ROC graph. The following corollary describes the cycle counts when the community size is a constant independent of d.

Corollary 33. *Let $G \sim ROC(n, d, \mu, 0)$ Then for $d = o\left(n^{\frac{1}{k-1}}\right)$,*

$$\mathsf{E}[C_k(G)] = c(k)nd + o(nd).$$

Proof. Let $P \in P_S$, and let $X_P(G)$ be the random variable for the number of walks in G with permutation type P. We show that $\mathsf{E}[X_P(G)]$ is the same for each $P \in P_S$ and so

$$\mathsf{E}[X_S(G)] = \sum_{P \in P_S} \mathsf{E}[X_P(G)] = |P_S| \, \mathsf{E}[X_P(G)]. \tag{14}$$

To compute the expectation of $X_P(G)$, we apply linearity of expectation to indicator random variables representing each possible walk. We define a possible walk as (i) an ordered set of vertices $(v_1, \ldots v_k)$ such that the walk $v_1, v_2, \ldots v_k$ is closed and has cycle permutation P and (ii) an ordered set of communities $(u_1, \ldots u_k)$, $u_i \in [xnd^{1-2a}]$. The walk exists if for each $1 \leq i \leq k-1$, the vertices v_i and v_{i+1} are adjacent by an edge that was added in the $(u_i)^{th}$ community in the construction of G. The probability a possible walk exists in G depends on how often the community labels $(u_1, \ldots u_k)$ change between adjacent vertices.

Let A be the set of possible walks in which each cycle is assigned a distinct community, each edge is labeled with the community assigned to its cycle, and there are $k - t + 1$ distinct vertices. We write $X_P(G) = A_P(G) + B_P(G)$ where $A_P(G)$ is the random variable for the number of walks in A that appear in G and $B_P(G)$ is the random variable for the number of walks that appear in G and are not in A. We compute $\mathsf{E}[A_P(G)]$ and show that $\mathsf{E}[B_P(G)] = o\left(\mathsf{E}[A_P(G)]\right)$ in cases (1) and (2) and $\mathsf{E}[B_P(G)] = \Theta\left(\mathsf{E}[A_P(G)]\right)$ in case (3).

Claim 1: $\mathsf{E}[A_P(G)] = \left(\prod_{i=1}^{j} c(a_i)^{t_i}\right) nd^{(1-2a)t+ak} + o\left(nd^{(1-2a)t+ak}\right)$.

We write $A_P(G)$ as the sum of random variables $A_W(G)$ that indicate if a walk $W \in A$ is in G. We show that $\mathsf{E}[A_W(G)]$ is the same for all W in A and so

$$\mathsf{E}[A_P(G)] = |A| \, \mathsf{E}[A_W(G)] = |A| \Pr[A_W(G)].$$

We now compute $\Pr[A_W(G)]$. Let $z_1, \ldots z_t$ be the non-zero characters of P ordered by first appearance. Let A_ℓ be the event that all edges in the cycle corresponding z_ℓ were added in the community assigned to z_ℓ, which we denote y_ℓ. The probability of A_ℓ depends on the community type (m_i, q_i, β_i) of y_ℓ. We compute

$$\mathsf{Pr}[A_\ell] = \sum_i \mathsf{Pr}[\text{ specified cycle appears in community } y_\ell \mid\ y_\ell \text{ is type } i]\,\mathsf{Pr}[\ y_\ell \text{ is type } i].$$

It follows that

$$\mathsf{Pr}[A_\ell] = \begin{cases} \sum_{i\in B} 2\left(\frac{m_i d^a}{n}\right)^{z_\ell} q_i^{z_\ell}\mu_i + \sum_{i\in B^c}\left(\frac{m_i d^a}{n}\right)^{z_\ell} q_i^{z_\ell}\mu_i & z_\ell \geq 3 \\ \\ \sum_{i\in B} 2\left(\frac{m_i d^a}{n}\right)^{2} q_i\mu_i + \sum_{i\in B^c}\left(\frac{m_i d^a}{n}\right)^{2} q_i\mu_i & z_\ell = 2. \end{cases}$$

Equivalently, $\mathsf{Pr}[A_\ell] = d^{az_\ell}c(z_\ell)/(xn^{z_\ell})$. The event that the walk W appears in G is the intersection of the events $A_1, \ldots, A_t$. Note that these events are independent because the communities $y_1, y_2, \ldots y_t$ are distinct. It follows

$$\mathsf{Pr}[A_W(G)] = \prod_{\ell=1}^{t}\mathsf{Pr}[A_\ell] = \frac{d^{ak}\prod_{i=1}^{j} c(a_i)^{t_i}}{n^k x^t}.$$

Next we compute the size of A. There are $\frac{(xnd^{1-2a})!}{(xnd^{1-2a}-t)!} = (xnd^{1-2a})^t + o\left((xnd^{1-2a})^t\right)$ ways to select t distinct communities and $\frac{n!}{n-(k-t+1)!} = n^{k-t+1} + o\left(n^{k-t+1}\right)$ ways to select the the vertices for $W \in A$. The claim follows,

$$\begin{aligned} \mathsf{E}[A_P(G)] &= \left((xnd^{1-2a})^t + o\left((xnd^{1-2a})^t\right)\right)\left(n^{k-t+1} + o\left(n^{k-t+1}\right)\right)\left(\frac{d^{ak}\prod_{i=1}^{j} c(a_i)^{t_i}}{n^k x^t}\right) \\ &= nd^{(1-2a)t+ak}\left(\prod_{i=1}^{j} c(a_i)^{t_i}\right) + o\left(nd^{(1-2a)t+ak}\right). \end{aligned} \tag{15}$$

<u>Claim 2:</u> In cases (1) and (2), $\mathsf{E}[B_P(G)] = o\left(nd^{(1-2a)t+ak}\right)$, and in case (3) $\mathsf{E}[B_P(G)] = \Theta\left(nd^{(1-2a)t+ak}\right)$.

Before computing $\mathsf{E}[B_P(G)]$, we introduce notation to describe the features of possible walks that are not in A. Let $z_1, \ldots, z_t$ be the non-zero characters of P, so the walk is composed of cycles of lengths $z_1, \ldots z_t$. Let m_ℓ be the number of vertices in the cycle corresponding z_ℓ that do not appear in the cycles corresponding to $z_1, \ldots z_{\ell-1}$. Let λ_i be the number of community edge labels in the cycle corresponding to z_ℓ that do not appear as community edge labels in any cycle corresponding to $z_1, \ldots z_{\ell-1}$. For each community assignment u_i, v_i and v_{i+1} must both be in community u_i if the possible walk exists in G. We say the i^{th} edge "assigns" the community u_i to the vertices v_i and v_{i+1}; if two consecutive edges have the same community label, then the common end is assigned to the same community twice. Let Γ_ℓ be the number of vertex-community assignments from the cycle corresponding to z_ℓ that are not assigned in cycles corresponding to $z_1, \ldots z_{\ell-1}$. Let $m = \sum_i m_i \leq k-t+1$, $\lambda = \sum_i \lambda \leq k$, $\Gamma = \sum \Gamma_i$, and j be the number of indices i such that $\lambda_i \geq 2$. Let $\mathcal{P}$ denote the parameters $\{\lambda_i, m_i, \Gamma_i\}$, and let $B_{\mathcal{P}}(G)$ be the number of possible walks with the parameters $\mathcal{P}$. There are $\Theta\left(\left(nd^{1-2a}\right)^\lambda\right)$ ways to select the communities, $\Theta\left(n^m\right)$ ways to select the vertices. The probability a vertex is an assigned community is $\Theta\left(\frac{d^a}{n}\right)$. It follows that

$$\mathsf{E}[W_{\mathcal{P}}(G)] = \Theta\left(\left(nd^{1-2a}\right)^{\lambda} n^m \left(\frac{d^a}{n}\right)^{\Gamma}\right). \tag{16}$$

Next we show that for any set of parameters $\mathcal{P}$, $\Gamma \geq m+\lambda+j-1$. First we describe relationships between λ_ℓ, Γ_ℓ, z_ℓ, and m_ℓ in different settings.

1. If there are precisely λ_ℓ community labels in the cycle corresponding to z_ℓ then

$$\begin{aligned} &\Gamma_\ell \geq z_\ell + \lambda_\ell && \lambda_\ell \geq 2 \\ &\Gamma_\ell = z_\ell && \lambda_\ell = 1. \end{aligned}$$

   If there are $\lambda_\ell \geq 2$ communities assigned to edges in the cycle, then there are at least λ_ℓ vertices where the adjacent edges are assigned different communities. These vertices contribute $2\lambda_\ell$ vertex-community assignments and the remaining $z_\ell - \lambda_\ell$ vertices are also assigned a community. If $\lambda_\ell = 1$, then each vertex is assigned to the one community.

2. If there are more than λ_ℓ community labels in the cycle corresponding to z_ℓ then

$$\begin{aligned} &\Gamma_\ell \geq m_\ell + \lambda_\ell + 1 && \lambda_\ell \geq 1 \\ &\Gamma_\ell \geq m_\ell && \lambda_\ell = 0. \end{aligned}$$

   The m_i new vertices must be assigned at least one community. When $\lambda_\ell \neq 0$, there must be at least $\lambda_\ell + 1$ vertices in which (i) both adjacent edges are labeled with two different first appearing communities or (ii) one adjacent edge is labeled with a first appearing community and one adjacent edge is labeled with a community that has already appeared. If such a vertex is a new vertex then this vertex has a total of two community assignments. If such a vertex has appeared before, it has not been previously assigned to a first appearing community, so this contributes one community assignment.

Since the first vertex of a cycle corresponding to z_ℓ for $\ell \geq 2$ has already been visited, $z_\ell \geq m_\ell + 1$ for $\ell \geq 2$. Therefore when $\ell \geq 2$

$$\begin{aligned} &\Gamma_\ell \geq m_\ell + \lambda_\ell + 1 && \lambda_\ell \geq 2 \\ &\Gamma_\ell = m_\ell + \lambda_\ell && \lambda_\ell \leq 1, \end{aligned}$$

and for $\ell = 1$,

$$\begin{aligned} &\Gamma_1 \geq m_1 + \lambda_1 && \lambda_1 \geq 2 \\ &\Gamma_1 = m_1 && \lambda_1 = 1. \end{aligned}$$

Summing the above inequalities over ℓ yields the observation that $\Gamma \geq m + \lambda + j - 1$. Note also that $m \leq k - t + 1$. Equation (16) becomes

$$\mathsf{E}[W_{\mathcal{P}}(G)] = \Theta\left(\left(nd^{1-2a}\right)^{\lambda} n^m \left(\frac{d^a}{n}\right)^{m+\lambda+j-1}\right) = \Theta\left(n^{1-j} d^{(1-a)\lambda + a(k-t+j)}\right). \tag{17}$$

Next consider a walk that is not in A. There must either be (i) a cycle that has at least two new community labels and so $j \geq 1$ or (ii) fewer than t total community labels, so $\lambda \leq t-1$. If $t = 1$ (the walk is a simple cycle), case (ii) does not occur because there must be at least one community label. In case (i), $j \geq 1$, $\lambda \leq k - t + j$ and $t \geq 1$, and so Equation (17) becomes

$$\mathsf{E}[W_{\mathcal{P}}(G)] = \Theta\left(n^{1-j} d^{k-t+j}\right) = \begin{cases} \Theta\left(d^{k-t}\right) & j = 1 \\ o\left(d^{k-t}\right) & j \geq 2. \end{cases} \tag{18}$$

Equivalently if $\mathcal{P}$ is type (i), then

$$\mathsf{E}[W_{\mathcal{P}}(G)] = nd^{(1-2a)t+ak} O\left(n^{-1} d^{(1-a)k+2a-1}\right) = o\left(nd^{(1-2a)t+ak}\right).$$

In case (ii), then $\lambda \leq t-1$ and $j \geq 0$, and so Equation (17) becomes

$$\mathsf{E}[W_{\mathcal{P}}(G)] = \Theta\left(n^{1-j} d^{(1-a)\lambda + a(k-t+j)}\right) = \begin{cases} \Theta\left(nd^{(1-2a)t+ak+a-1}\right) & \lambda = t-1 \text{ and } j = 0 \\ o\left(nd^{(1-2a)t+ak+a-1}\right) & \lambda \leq t-2 \text{ or } j \geq 1. \end{cases} \tag{19}$$

Equivalently if $\mathcal{P}$ is type (ii), then

$$\mathsf{E}[W_{\mathcal{P}}(G)] = O\left(n^{1-j} d^{(1-a)(t-1)+a(k-t+j)}\right) = nd^{(1-2a)t+ak} O\left(d^{a-1}\right).$$

Therefore for any $\mathcal{P}$ that is not in A,

$$\mathsf{E}[W_{\mathcal{P}}(G)] = \begin{cases} o\left(nd^{(1-2a)t+ak}\right) & a < 1 \text{ or } a = 1 \text{ and } t = 1 \\ \Theta\left(nd^{(1-2a)t+ak}\right) & a = 1 \text{ and } t > 1. \end{cases}$$

Note $\mathsf{E}[B_P(G)] = \sum_{\mathcal{P}} \mathsf{E}[W_{\mathcal{P}}(G)]$. Since the number of sets of valid parameters $\mathcal{P}$ is constant, claim 2 follows.

The computation of $\mathsf{E}[X_P(G)] = \mathsf{E}[A_P(G)] + \mathsf{E}[B_P(G)]$ did not rely on any information about P besides that $P \in P_S$. Therefore, equation (14) holds and the statement of the lemma follows directly from claims 1 and 2. □

Proof. (of Theorem 27) For $a < 1$, Lemma 32 implies

$$\mathsf{E}[W_k(G)] = \sum_{S \in \mathcal{S}} |P_s| \left(\prod_{i=1}^{j} c(a_i)^{t_i}\right) nd^{(1-2a)t+ak} + o\left(nd^{(1-2a)t+ak}\right) \tag{20}$$

We now collect the highest order terms of (20) for different values of a. Recall $\sum a_i t_i = k$ and $a_i \geq 2$, so $1 \leq \sum t_i \leq \frac{k}{2}$.

Case 1: $a \in (0, 1/2)$**.** The highest order term of (20) is from $S \in \mathcal{S}$ with $a_1 = 2$ and $t_1 = \frac{k}{2}$ for k even and $S \in \mathcal{S}$ with $a_1 = 3$, $a_2 = 2$, $t_1 = 1$, $t_2 = \frac{k-3}{2}$ for k odd. For even k, Equation (20) becomes

$$\mathsf{E}[W_k(G)] = nd^{k/2} \frac{k!}{\left(\frac{k}{2}\right)! \left(\frac{k}{2}+1\right)!} c(2)^{k/2} + o\left(nd^{k/2}\right) = (Cat_{k/2}) nd^{k/2} + o(nd^{k/2}).$$

For odd k, Equation (20) becomes

$$\mathsf{E}[W_k(G)] = O\left(nd^{ak+(1-2a)\frac{k-1}{2}}\right) = O\left(nd^{\frac{k-1}{2}+a}\right) = o(nd^{k/2}).$$

It follows that the ROC family (μ, a) achieves the Catalan vector with sparsity exponent 1/2, and achieves the $k-2$ length prefix with k-sparsity exponent 1/2 for all $k \geq 4$.

Case 2: $a = 1/2$**.** Each $S \in \mathcal{S}$ contributes a term of order $d^{k/2}$ to Equation (20). Therefore

$$\mathsf{E}[W_k(G)] = \left(\sum_{S\in\mathcal{S}_k} \frac{k!}{(\prod t_i!)(k+1-\sum t_i)!} \prod_i c(a_i)^{t_i}\right) nd^{k/2} + o\left(nd^{k/2}\right) = w(k)nd^{k/2} + o\left(nd^{k/2}\right).$$

It follows that the ROC family (μ, a) achieves the limit $(w_3, w_4, \dots)$ with sparsity exponent 1/2, and achieves the $k-2$ length prefix with k-sparsity exponent 1/2 for all $k \geq 3$.

Case 3: $a \in (1/2, 1)$**.** The highest order term of (20) is from $S \in \mathcal{S}$ with $a_1 = k$ and $t_1 = 1$. Therefore Equation (20) becomes

$$\mathsf{E}[W_k(G)] = c(k)nd^{1+a(k-2)} + o\left(nd^{1+a(k-2)}\right).$$

It follows that the ROC family (μ, a) achieves the limit $(c_3, c_4, \dots)$ with sparsity exponent a, and achieves the $k-2$ length prefix with k-sparsity exponent a for all $k \geq 3$.

Case 4: $a = 1$**.** For $S \in \mathcal{S}$ with $t = \sum t_i$, the number of walks with permutation type in the set P_S is $\Theta\left(nd^{(1-2a)t+ak}\right)$. Therefore the walks contributing the highest order terms correspond to $S \in \mathcal{S}$ with $a_1 = k$ and $t_1 = 1$. By parts 2 and 3 of Lemma 32, we have

$$\mathsf{E}[W_k(G)] = c(k)nd^{k-1} + o\left(nd^{k-1}\right).$$

It follows that the ROC family (μ, a) achieves the limit $(c_3, c_4, \dots)$ with sparsity exponent 1, and achieves the $k-2$ length prefix with k-sparsity exponent 1 for all $k \geq 3$. □

4.2.4 The convergence of sequences of ROC graphs

Definition 5 states that the vector achieved by a ROC family is the expected walk count of a ROC graph from that family normalized with respect to *expected* degree. We now justify this definition by showing that for $G \sim ROC(n, d, \mu, a)$, the probability that the normalized closed walk count $W_j(G, \alpha)$ deviates from the limit w_j achieved by the family tends to zero as d grows (Theorem 34). Moreover, we show that the sequence $G_i \sim ROC(n_i, d_i, \mu, a)$ almost surely converges to the limit achieved by the family when n_i and d_i grow sufficiently fast (Corollary 35).

Theorem 34. *Let $(w_3, w_4, \dots)$ be the limit achieved by the ROC family (μ, a). Let $G \sim ROC(n, d, \mu, a)$ where $d = o\left(n^{1/((1-a)k+2a-1)}\right)$ and $|w_j - \frac{\mathsf{E}[W_j(G)]}{nd^{1+\alpha(j-2)}}| < \varepsilon/2$. Then for $\alpha = \max\{a, 1/2\}$,*

$$\Pr[|W_j(G, \alpha) - w_j| > \varepsilon] = f(d, a)$$

where

$$f(d,a) = \begin{cases} O\left(d^{-1+2a} + \frac{d^{k/2-1}}{n} + \frac{d^{(k-1)(2a-1)}}{n}\right) & a < 1/2 \\ O\left(\frac{d^{k/2-1}}{n} + d^{-1/2}\right) & a = 1/2 \\ O\left(d^{1-2a} + \frac{d^{(1-a)(k-2)}}{n}\right) & 1/2 < a < 1 \\ O\left(d^{-1} + \frac{d}{n}\right) & a = 1. \end{cases}$$

Corollary 35. *Let $(w_3, w_4, \dots)$ be the limit achieved by the ROC family (μ, a). Let $G_i \sim ROC(n_i, d_i, \mu, a)$ where $d_i = o\left(n_i^{1/((1-a)k+1-2a)}\right)$ and $f(d_i, a)$ is defined for G_i as in Theorem 34. If $\sum_{i=1}^{\infty} f(d_i, a) < \infty$ the sequence of graphs (G_i) converges to the limit $(w_3, w_4, \dots)$ with sparsity exponent $\alpha = \max\{a, 1/2\}$.*

Achieving normalized and unnormalized closed walk counts. A ROC family (μ, a) that achieves the limit of a sequence of graphs (G_i) with the appropriate sparsity exponent is a sampleable model that produces graphs in which the normalized closed walk counts match the limit up to an error term that tends to zero as the size of the sampled graph grows. The following remark describes when a sequence of graphs drawn from the ROC model also matches the unnormalized closed walk counts of the sequence (G_i) term by term. The remark is stated for k-convergence and k-limits, but an analogous statement holds for full convergence and limits.

Remark 36. *Let (G_i) be a sequence of graphs each with n_i vertices and average degree d_i such that (G_i) is k-convergent with k-limit L and k-sparsity exponent α. Suppose the ROC family (μ, a) achieves the limit L with sparsity exponent α.*

1. *If $d_i = o\left(n_i^{1/((1-a)k+2a-1)}\right)$ then the sequence (H_i) with $H_i \sim ROC(n_i, d_i, \mu, a)$ has the property that for sufficiently large i and $j \leq k$, in expectation G_i and H_i have the same average degree and number of closed j-walks up to lower order terms with respect to d_i.*
2. *It is possible to construct other sequences (H_i) with $H_i \sim ROC(n_i, f(n_i), \mu, a)$ such that for sufficiently large i, in expectation H_i and G_i have different edge densities, but have the same normalized number of closed walks up to lower order terms.*

To prove Theorem 34, we will apply Lemma 23 , which bounds the probability that the normalized walk count deviates from expectation in terms of the probability that the number of edges deviates and the probability that the walk count deviates. Lemmas 37 and 38 compute these quantities.

Lemma 37. *Let $G \sim ROC(n, d, \mu, a)$. Let $D(G)$ be the random variable for the average degree of G. Then*

$$\mathsf{E}[D(G)] = d \quad \text{and} \quad \mathsf{Var}[D(G)] = \Theta\left(\frac{d^{1+a}}{n}\right).$$

Proof. Let $D(G) = \frac{1}{n}\sum_{v,w,u} X_{u,v,w}$ where $X_{u,v,w}$ is an indicator random variable for the event that the edge w, v is added in the u^{th} community. Note

$$\mathsf{E}[X_{u,v,w}] = \mathsf{Pr}[X_{u,v,w}] = \sum_{i \in B^c} \mu_i \left(\frac{m_i d^a}{n}\right)^2 q_i + \sum_{i \in B} \mu_i 2 \left(\frac{m_i d^a}{n}\right)^2 q_i = \frac{d^{2a}}{xn^2}.$$

There are $n(n-1)$ pairs w, v and xnd^{1-2a} communities u. Therefore

$$\mathsf{E}[D(G)] = \frac{1}{n}\sum_{v,w,u} \mathsf{E}[X_{u,v,w}] = \frac{1}{n} n(n-1)xnd^{1-2a}\frac{d^{2a}}{xn^2} = d - \frac{d}{n}.$$

Next we compute the expected pairs of edges, $\mathsf{E}\left[D(G)^2\right]$. We fix the potential edge defined by vertices a and b and community x and sum over all other potential edges.

$$
\begin{aligned}
\mathsf{E}\left[D(G)^2\right] &= \left(\frac{1}{n^2}\right) n(n-1)(xnd^{1-2a}) \sum_{u,v,w} \mathsf{Pr}[X_{u,v,w} \text{ and } X_{a,b,x}] \\
&= x(n-1)d^{1-2a}\Bigg(\sum_{u,v,w\neq x} \mathsf{Pr}[X_{u,v,w}]\,\mathsf{Pr}[X_{a,b,x}] + \sum_{u,v\notin\{a,b\},x} \mathsf{Pr}[X_{u,v,w}]\,\mathsf{Pr}[X_{a,b,x}] \\
&\quad + 2\sum_{u=a,v\neq b,x} \mathsf{Pr}[X_{u,v,w} \text{ and } X_{a,b,x}] + \mathsf{Pr}[X_{a,b,x}]\Bigg) \\
&= x(n-1)d^{1-2a}\left(\frac{d^{2a}}{xn^2}\right)\left(n(n-1)\left(xnd^{1-2a}-1\right)\left(\frac{d^{2a}}{xn^2}\right) + (n-2)(n-3)\left(\frac{d^{2a}}{xn^2}\right)\right. \\
&\quad \left. + (n-3)\Theta\left(\frac{d^a}{n}\right) + 1\right) \\
&= \left(d - \frac{d}{n}\right)^2 + \Theta\left(\frac{d^{1+a}}{n}\right).
\end{aligned}
$$

It follows that

$$
\mathsf{Var}[D(G)] = \mathsf{E}\left[D(G)^2\right] - \mathsf{E}[D(G)]^2 = \Theta\left(\frac{d^{1+a}}{n}\right).
$$

□

Lemma 38. *Let $G \sim ROC(n,d,\mu,a)$ with $d = o\left(n^{1/((1-a)k+2a-1)}\right)$. Let $W_k(G)$ be the random variable for the number of closed k-walks in G. Then*

$$
\mathsf{Var}[W_k(G)] = \begin{cases}
O\left(\left(nd^{k/2}\right)^2\left(d^{-1+2a} + \frac{d^{k/2-1}}{n} + \frac{d^{(k-1)(2a-1)}}{n}\right)\right) & a < 1/2 \\
O\left(\left(nd^{k/2}\right)^2\left(\frac{d^{k/2-1}}{n} + d^{-1/2}\right)\right) & a = 1/2 \\
O\left(\left(nd^{1+a(k-2)}\right)^2\left(d^{1-2a} + \frac{d^{(1-a)(k-2)}}{n}\right)\right) & 1/2 < a < 1 \\
O\left(\left(nd^{k-1}\right)^2\left(d^{-1} + \frac{d}{n}\right)\right) & a = 1.
\end{cases}
$$

Proof. We give an upper bound on $\mathsf{E}\left[W_k(G)^2\right]$ by counting the expected number of pairs of walks. Let $P_k'(G)$ be the random variable for the number of pairs of k-walks in G that do not intersect, and let $P_k''(G)$ be the random variable for the number of pairs of k-walks in G that do intersect. Note that two k-walks that intersect can be thought of as a $2k$ walk. The expected number of $2k$ walks in G is $\Theta\left(nd^{1+a(2k-2)}\right)$ (see Theorem 27), and so $\mathsf{E}[P_k''(G)] = \Theta\left(nd^{1+a(2k-2)}\right)$.

To compute $P_k'(G)$ we recall the partition of possible walks with permutation type P_S into sets $A, B(i)$ and $B(ii)$ as described in the proof of Lemma 32. Let $S(t)$ be the set of $S \in \mathcal{S}_k$ such that $\sum t_i = t$. For $S \in S(t)$, the expected number of walks with type A is at most $nd^{(1-2a)t+ak}\left(\prod_{i=1}^{j} c(a_i)^{t_i}\right)$ (see (15)). The expected number of walks with type $B(i)$ is $\Theta\left(d^{k-t}\right)$ (see (18)), and the expected number of walks with type $B(ii)$ is $\Theta\left(nd^{(1-2a)t+ak+a-1}\right)$ when $t \neq 1$ and 0 when $t = 1$ (see (19)). Therefore

$$
\mathsf{E}\left[P_k'(G)\right] \leq \left(\sum_{t=1}^{\lfloor k/2\rfloor} \sum_{S\in S(t)} |P_S| nd^{(1-2a)t+ak}\left(\prod_{i=1}^{j} c(a_i)^{t_i}\right) + \Theta\left(d^{k-t} + \zeta_t nd^{(1-2a)t+ak+a-1}\right)\right)^2
$$

where $\zeta_t = 0$ if $t = 1$ and $\zeta_t = 1$ otherwise. We simplify and obtain

$$\mathsf{E}\left[P_k'(G)\right] = \begin{cases} \mathsf{E}[W_k(G)]^2 + O\left(nd^{k/2}\left(nd^{k/2-1+2a} + d^{k-1} + nd^{k/2+a-1}\right)\right) & a < 1/2 \\ \mathsf{E}[W_k(G)]^2 + O\left(nd^{k/2}\left(d^{k-1} + nd^{ak+a-1}\right)\right) & a = 1/2 \\ \mathsf{E}[W_k(G)]^2 + O\left(nd^{1+a(k-2)}\left(nd^{2+a(k-4)} + d^{k-1} + nd^{1+a(k-3)}\right)\right) & 1/2 < a < 1 \\ \mathsf{E}[W_k(G)]^2 + O\left(nd^{k-1}\left(nd^{k-2} + d^{k-1} + nd^{k-2}\right)\right) & a = 1. \end{cases}$$

Finally we compute

$$\begin{aligned} \mathsf{Var}[W_k(G)] &= \mathsf{E}\left[W_k(G)^2\right] + \mathsf{E}[W_k(G)]^2 = \mathsf{E}\left[P_k'(G)\right] + \mathsf{E}\left[P_k''(G)\right] - \mathsf{E}[W_k(G)]^2 \\ &= \begin{cases} O\left(nd^{k/2}\left(nd^{k/2-1+2a} + d^{k-1} + nd^{k/2+a-1}\right)\right) + O\left(nd^{1+a(2k-2)}\right) & a < 1/2 \\ O\left(nd^{k/2}\left(d^{k-1} + nd^{ak+a-1}\right)\right) + O\left(nd^{k}\right) & a = 1/2 \\ O\left(nd^{1+a(k-2)}\left(nd^{2+a(k-4)} + d^{k-1} + nd^{1+a(k-3)}\right)\right) + O\left(nd^{1+a(2k-2)}\right) & 1/2 < a < 1 \\ O\left(nd^{k-1}\left(nd^{k-2} + d^{k-1} + nd^{k-2}\right)\right) + O\left(nd^{2k-1}\right) & a = 1, \end{cases} \end{aligned}$$

and the statement follows by simplifying the above expressions. □

We now prove Theorem 34 by applying Lemma 23.

Proof. (of Theorem 34) Let $g = w_j - \frac{\mathsf{E}[W_j(G)]}{nd^{1+\alpha(j-2)}}$, $\delta = \min\left\{\frac{\varepsilon}{2j(w_j+\varepsilon)}, \frac{1}{2(j-1)^2}\right\}$ and $\lambda = \varepsilon/2 - |g|$. By Lemmas 23, 37 and 38,

$$\mathsf{Pr}[|W_j(G,\alpha) - w_j| > \varepsilon] \leq \frac{\mathsf{Var}[D(G)]}{\delta^2 d^2} + \frac{\mathsf{Var}[W_j(G)]}{\lambda^2\left(nd^{1+\alpha(j-2)}\right)^2} \leq O\left(\frac{d^{a-1}}{n}\right) + f(d,a) = f(d,a).$$

□

Corollary 35 follows directly from Theorem 34 and part 3 of Lemma 22.

4.3 Conditions for ROC achievable limits

In this section we address the questions: for which vectors L does there exist a ROC family that achieves limit or k-limit L with sparsity exponent α? We first show that all 4-limits are achievable in Section 4.3.1, then describe necessary and sufficient conditions for a limit vector (of any length) to be achievable in Section 4.3.2. Finally Lemma 46 in Section 4.3.3 gives a convenient criterion for determining when the Stietljes condition is satsified.

4.3.1 Achievability of (w_3, w_4)

In this section we prove Theorem 6, which states that any (w_3, w_4) that is a limit of a sequence of graphs with increasing degree can be achieved by a ROC family. In fact, the requirement that degree increases is only needed for the case in which the 4-sparsity exponent is 1/2.

Lemma 39. *Let $C_j(G)$ and $W_j(G)$ denote the number of simple j-cycles and closed j-walks of a graph G respectively. For any graph G on n vertices with average d*

$$W_4(G) \geq \frac{W_3(G)^2}{nd} \quad \text{and} \quad C_4(G) \geq C_3(G)\left(\frac{C_3(G)}{nd} - 1\right).$$

Proof. For each directed edge $e = (u, v)$ let t_e be the number of walks that traverse a triangle with first edge (u, v). For each edge (u, v) we can construct t_e^2 four walks including $\binom{t_e}{2}$ four cycles as follows. Select two triangles (u, v, a) and (u, v, b). The closed walk (u, b, v, a) is a closed four walk. When $a \neq b$ the walk is a four cycle. Note $C_3(G) = W_3(G)$. It follows that

$$W_4(G) \geq \sum_{e \in E(G)} t_e^2 \geq nd \left(\frac{W_3(G)}{nd}\right)^2 \quad \text{and} \quad C_4(G) \geq \sum_{e \in E(G)} \binom{t_e}{2} \geq nd \left(\frac{C_3(G)}{nd}\right)\left(\frac{C_3(G)}{nd} - 1\right).$$

□

Note the first part of the lemma also follows from the observation that for any set of λ_i,

$$\left(\sum_i \lambda_i^2\right)\left(\sum_i \lambda_i^4\right) \geq \left(\sum_i \lambda_i^3\right)^2.$$

Using these properties, we now prove Theorem 6.

Proof. (of Theorem 6) Let α be the 4-sparsity exponent of the sequence.

Case 1: $\alpha > 1/2$. By Lemma 39 for each graph G_i in the sequence satisfies

$$W_4(G_i, \alpha) \geq W_3(G_i, \alpha)^2.$$

It follows that $w_4 \geq w_3$. If $w_3 \neq 0$, the ROC family (μ, a) where μ is the distribution with support one on $m = w_4^2/w_3^3$ and $q = w_3^2/w_4$ achieves the limit (w_3, w_4). If $w_3 = 0$, the bROC family (μ, a) where μ is the distribution with support one on $m = w_4$ and $q = 1$ achieves the limit (w_3, w_4).

Case 2: $\alpha = 1/2$. It suffices to show that the cycle counts $T((w_3, w_4)) = (w_3, w_4 - 2)$ are the moments of some distribution. By Lemma 39, $C_4(G) \geq C_3(G)\left(\frac{C_3(G)}{nd} - 1\right) = W_3(G)\left(\frac{W_3(G)}{nd} - 1\right)$. Let $T_4(G)$ be the number of closed four walks that trace a path of length two. The number of two paths is $\sum_v \binom{deg(v)}{2} \geq n\binom{d}{2}$, and each two path contributes four closed four walks. Each edge contributes two closed four walks. It follows that

$$W_4(G) = C_4(G) + 4T_4(G) + nd \geq W_3(G)\left(\frac{W_3(G)}{nd} - 1\right) + 2nd(d-1) + nd,$$

and so

$$W_4(G_i, 1/2) \geq W_3(G_i, 1/2)^2 + 2 + O\left(\frac{1}{\sqrt{d}}\right).$$

Therefore $w_4 \geq w_3^2 + 2$. If $w_3 \neq 0$, the ROC family (μ, a) where μ is the distribution with support one on $m = (w_4 - 2)^2/w_3^3$ and $q = w_3^2/(w_4 - 2)$ achieves the limit (w_3, w_4). If $w_3 = 0$, the bROC family (μ, a) where μ is the distribution with support one on $m = w_4 - 2$ and $q = 1$ achieves the limit (w_3, w_4). □

4.3.2 Achievability of limits of general sequences

In this section we prove Theorems 4 and 5, which characterize achievable k-limits for sparsity exponent greater than half and half respectively. Additionally, we prove the analogous characterization for full achievability, as stated in the following theorems.

Theorem 40 (full achievability with sparsity exponent $> 1/2$). *A limit vector $(w_3, w_4, \dots)$ is achievable by ROC with sparsity exponent greater than $1/2$ if and only if there exists $\gamma \in [0,1]$, $s_0, s_1, \dots, t_0, t_2, \dots \in \mathbb{R}^+$, $s_2, t_2 \leq 1$ such that $(s_0, s_1, s_2, \dots)$ and $(t_0, t_2, \dots)$ satisfy the full Stieltjes condition and for all $j \geq 3$*

$$w_j = \begin{cases} \gamma s_j & j \text{ odd} \\ \gamma s_j + (1-\gamma) t_j & j \text{ even.} \end{cases}$$

Theorem 41 (full achievability with sparsity exponent $1/2$). *Let $T((c_3, c_4, \dots c_k)) = (w_3, w_4, \dots w_k)$ be the transformation of a vector given in Definition 8. The limit vector $(w_3, w_4, \dots w_k)$ is achievable by ROC with sparsity exponent $1/2$ if and only if there exists $\gamma \in [0,1]$, $s_0, s_1, \dots, t_0, t_2, \dots \in \mathbb{R}^+$, $s_2, t_2 \leq 1$ such that $(s_0, s_1, s_2, \dots)$ and $(t_0, t_2, \dots)$ satisfy the full Stieltjes condition and for all $j \geq 3$*

$$c_j = \begin{cases} \gamma s_j & j \text{ odd} \\ \gamma s_j + (1-\gamma) t_j & j \text{ even.} \end{cases}$$

The question underlying achievability is how to determine when a vector is the vector of normalized cycle counts of some ROC family. Note that the normalized cycle counts $(c(3), c(4), \dots c(k))$ of the family $ROC(n, d, \mu, a)$ are the moments of a discrete probability distribution over values determined by m_i, q_i and β_i scaled by x. The question of whether a vector can be realized as the vector of normalized cycle counts for some ROC family is a slight variant of the Stieltjes moment problem, which gives necessary and sufficient conditions for a sequence to be the moment sequence of some distribution with positive support.

Our question differs in two key ways. First, the second moment is not directly specified; instead we obtain an upper bound on the second moment from the restriction that

$$x \left(\sum_{i \in B^c} \mu_i m_i^2 p_i + 2 \sum_{i \in B} \mu_i m_i^2 p_i \right) = 1.$$

Second, for achievability of k-limits we are interested in when a vector is the prefix of some moment sequence.

The proofs of Theorems 40 and 41 rely on the classical solution to the Stieltjes moment problem (Lemma 42), and the proofs of Theorems 4 and 5 use a variant for truncated moment vectors (Lemma 43). We use these lemmas to show Lemma 44 and Lemma 45 which together with Theorem 27 directly imply the necessary and sufficient conditions given in Theorems 4, 5, 40 and 41. Finally we prove Lemma 46 which gives a sufficient local condition to guarantee that a sequence can be extended to satisfy the Stieltjes condition. The proof of this lemma establishes the semi-definiteness of Hankel matrices of sequences satisfying a logconcavity condition.

Lemma 42 (Stieltjes moment problem, see [34]). *A sequence $\mu = (\mu_0, \mu_1, \mu_2, \dots)$ is the moment sequence of a distribution with finite positive support if there exists $\{(x_i, t_i)\}$ with $x_i, t_i > 0$ such that $\sum_{i=1}^k x_i t_i^\ell = \mu_\ell$. A sequence μ is a moment sequence with positive support of size k if and only if*

$$\Delta_i^{(0)} > 0 \text{ and } \Delta_i^{(1)} > 0 \text{ for all } i < k \quad \text{and} \quad \Delta_i^{(0)} = \Delta_i^{(1)} = 0 \text{ for all } i \geq k$$

where Δ is given in Definition 6.

The next lemma follows, we give a proof for convenience.

Lemma 43 (truncated Stieltjes moment problem)**.** *A vector $\mu = (\mu_0, \mu_1, \mu_2, \dots, \mu_n)$ is the truncated moment sequence of a distribution with finite positive support if there exists $\{(x_i, t_i)\}$ with $x_i, t_i > 0$ such that $\sum_{i=1}^{k} x_i t_i^\ell = \mu_\ell$ for all $\ell \le n$. A vector μ is a truncated moment sequence with finite positive support if and only if the Stietljes condition given in Definition 6 is satisfied.*

Proof. Lemma 42 directly implies that a truncated moment vector μ satisfies the Stieltjes condition. To show that a vector μ satisfying the Stieltjes condition is a truncated moment vector with finite positive support, it suffices to construct $\mu_{n+1}, \mu_{n+2}, \dots$ such that the complete vector satisfies the hypotheses of Lemma 42. Given μ_{2s}, take μ_{2s+1} to be a value that makes $\Delta_s^{(1)} = 0$. Given μ_{2s+1}, take μ_{2s+2} to be the value that makes $\Delta_{s+1}^{(0)} = 0$. □

Lemma 44. *There exists s_0, s_1, s_2 with $s_2 \le 1$ such that $(s_0, s_1, s_2, a_3, \dots a_n)$ satisfies the Stieltjes condition if and only if there exists x_i, m_i, q_i with $x_i, m_i > 0$ and $0 \le q_i \le 1$ satisfying*

1. $\sum x_i m_i^2 q_i = 1$
2. $\sum x_i (m_i q_i)^j = a_j \quad$ *for all $3 \le j \le n$.*

Similarly, there exists s_0, s_1, s_2 with $s_2 \le 1$ such that $(s_0, s_1, s_2, a_3, \dots)$ satisfies the full Stieltjes condition if and only if there exists x_i, m_i, q_i with $x_i, m_i > 0$ and $0 \le q_i \le 1$ satisfying (1) and (2) for all $j \ge 3$.

Proof. First assume $(s_0, s_1, s_2, a_3, \dots a_n)$ satisfies the Stieltjes condition (or $(s_0, s_1, s_2, a_3, \dots)$ satisfies the full Stieltjes condition) and $s_2 \le 1$. By Lemma 43 (or Lemma 42) there exists a discrete distribution on $(t_1, t_2, \dots t_k)$ where t_i has mass x_i, $t_i > 0$, and

$$\sum x_i t_i^j = \begin{cases} a_j & 3 \le j \le n \text{ (or } j \ge 3) \\ s_i & 0 \le j \le 2. \end{cases}$$

Let $q_i = s_2$ for all i, and $m_i = t_i / s_2$ for all i. Observe

$$\sum x_i (m_i q_i)^j = \sum x_i \left(\frac{t_i s_2}{s_2} \right)^j = a_j \text{ for all } 3 \le j \le n \text{ (or for all } j \ge 3)$$

$$\sum x_i m_i^2 q_i = \sum \frac{x_i t_i^2}{s_2} = 1.$$

Next assume there exists x_i, m_i, q_i satisfying the given conditions. Let $s_j = \sum x_i t_i^j$ for $j \in \{1, 2, 3\}$ and $t_i = m_i q_i$. Note

$$\sum \alpha_i t_i^j = \sum x_i (m_i q_i)^j = a_j \text{ for all } 3 \le j \le n \text{ (or for all } j \ge 3),$$

and so $(s_0, s_1, s_2, a_3, \dots a_n)$ (or $(s_0, s_1, s_2, a_3, \dots)$) is a moment vector of a finite distribution with positive support. It follows by Lemma 43 (or Lemma 42) that the moment vector satisfies the (full) Stieltjes condition. To see that $s_2 \le 1$, let $q = \max_i q_i$ and observe

$$s_2 = \sum x_i m_i^2 q_i^2 \le q \sum x_i m_i^2 q_i = q \le 1.$$

□

Lemma 45. *There exists s_0, s_2 with $s_2 \leq 1$ such that $(s_0, s_2, a_4, \ldots a_n)$ satisfies the Stieltjes condition if and only if there exists x_i, m_i, q_i with $x_i, m_i > 0$ and $0 \leq q_i \leq 1$ satisfying*

1. $2 \sum x_i m_i^2 q_i = 1$
2. $2 \sum x_i (m_i q_i)^{2j} = a_{2j}$ *for all* $2 \leq j \leq n$.

Similarly, there exists s_0, s_2 with $s_2 \leq 1$ such that $(s_0, s_2, a_4, \ldots a_n)$ satisfies the full Stieltjes condition if and only if there exists x_i, m_i, q_i with $x_i, m_i > 0$ and $0 \leq q_i \leq 1$ satisfying (1) and (2) for all $j \geq 4$.

Proof. First assume $(s_0, s_2, a_4, a_6, \ldots a_n)$ satisfies the Stieltjes condition (or $(s_0, s_2, a_4, a_6, \ldots)$ satisfies the full Stieltjes condition) and $s_2 \leq 1$. It follows that $\left(\frac{s_0}{2}, \frac{s_2}{2}, \frac{a_4}{2}, \frac{a_6}{2}, \ldots \frac{a_n}{2}\right)$ satisfies the Stieltjes condition (or $\left(\frac{s_0}{2}, \frac{s_2}{2}, \frac{a_4}{2}, \frac{a_6}{2}, \ldots\right)$ satisfies the full Stieltjes condition) because multiplying all entries of a matrix by a positive number does not change the sign of the determinant. By Lemma 43 (or Lemma 42) there exists a discrete distribution on $(t_1, t_2, \ldots t_k)$ where t_i has mass x_i, $t_i > 0$, and

$$2 \sum x_i t_i^j = \begin{cases} a_{2j} & 2 \leq j \leq n \text{ (or } j \geq 4t) \\ s_{2j} & 0 \leq j \leq 1. \end{cases}$$

Let $q_i = s_2$, and $m_i = \sqrt{t_i}/s_2$ for all i. Observe

$$2 \sum x_i (m_i q_i)^{2j} = \sum x_i t_i^j = a_{2j} \text{ for all } 2 \leq j \leq n \text{ (or } j \geq 4)$$

$$2 \sum x_i m_i^2 q_i = \frac{2}{s_2} \sum x_i t_i = 1.$$

Next assume there exists x_i, m_i, q_i satisfying the given conditions. Let $s_0 = 2 \sum x_i$ and $t_i = (m_i q_i)^2$ for all i. Note

$$\sum x_i t_i^j = \sum \frac{x_i}{s_0} (m_i q_i)^{2j} = \frac{a_j}{2} \text{ for all } 2 \leq j \leq n \text{ (or } j \geq 4).$$

Let $q = \max_i q_i$, $s_1 = 2 \sum x_i t_i$, and observe

$$s_1 = 2 \sum x_i t_i = 2 \sum x_i m_i^2 q_i^2 \leq 2q \sum x_i m_i^2 q_i = q \leq 1.$$

It follows that $\left(\frac{s_0}{2}, \frac{s_2}{2}, \frac{a_4}{2}, \ldots \frac{a_n}{2}\right)$ is a moment vector (or $\left(\frac{s_0}{2}, \frac{s_2}{2}, \frac{a_4}{2}, \ldots\right)$ is a moment vector), and therefore by Lemma 43 (or Lemma 43) satisfies the (full) Stieltjes condition. It follows that $(s_0, s_1, a_2, \ldots a_n)$ also satisfies the Stieltjes condition (or $(s_0, s_1, a_2, \ldots)$ also satisfies the full Stieltjes condition) because multiplying all entries of a matrix by a positive number does not change the sign of the determinant. $\square$

Proof. (of Theorems 4, 5, 40 and 41) First assume the vectors of s_i and t_i satisfy the hypotheses. Then by Lemma 44, there exists (x_i, m_i, q_i) satisfying $\sum x_i (m_i q_i)^j = s_j$ and $\sum x_i m_i^2 q_i = 1$. For each triple add the triple $(m_i, q_i, \beta_i = 0)$ to the distribution. Soon we will specify the corresponding probability μ_i. By Lemma 45, there exists (x_i, m_i, q_i) satisfying $2 \sum x_i (m_i q_i)^{2j} = t_{2j}$ and $2 \sum x_i m_i^2 q_i = 1$. For each triple add the triple $(m_i, q_i, \beta_i = 1)$ to the distribution. Let B be the set of indices i such that $\beta_i = 1$ and B^c be the set of indices i such that $\beta_i = 0$. Let

$z = \sum_{i\in B} x_i\gamma + \sum_{i\in B^c} x_i(1-\gamma)$. We now define μ by assigning probabilities to triples (m_i, q_i, β_i). If $i \in B^c$, let $\mu_i = x_i\gamma/z$. If $i \in B$, let $\mu_i = x_i(1-\gamma)/z$. Note $\sum \mu_i = 1$, and therefore (μ, a) is a well-defined ROC family with

$$x = 1/\left(\sum_{i\in B^c} \mu_i m_i^2 q_i + 2\sum_{i\in B} \mu_i m_i^2 q_i\right) = z.$$

Theorem 27 implies that the family achieves the desired limit with sparsity exponent a.

Suppose the limit is achievable by some ROC family (μ, a). Let $\gamma = x\sum_{i\in B^c} \mu_i m_i^2 q_i$, and so $1-\gamma = 2x\sum_{i\in B} \mu_i m_i^2 q_i$. For each $i \in B^c$, let $x_i = x\mu_i/\gamma$. Note $\sum_{i\in B^c} x_i m_i^2 q_i = 1$, and so by Lemma 44, the vector with $s_j = \sum_{i\in B^c} \mu_i(m_i q_i)^j$ satisfies the Stietljes condition. For each $i \in B$, let $x_i = x\mu_i/(1-\gamma)$. Note $2\sum_{i\in B} x_i m_i^2 q_i = 1$, and so by Lemma 45, the vector with $t_{2j} = \sum_{i\in B} \mu_i(m_i q_i)^{2j}$ satisfies the Stietljes condition. Theorem 27 implies that c_j or w_j is the appropriate combination of s_j and t_j. □

Finally we show that a similar argument proves the condition for when it is possible to match a k-cycle-to-edge vector with a ROC family.

Proof. (of Theorem 8) First assume the vectors of s_i and t_i satisfy the hypotheses. Then by Lemma 44, there exists (x_i, m_i, q_i) satisfying $\sum x_i(m_i q_i)^j = s_j$ and $\sum x_i m_i^2 q_i = 1$. For each triple add the triple $(m_i, q_i, \beta_i = 0)$ to the distribution. Soon we will specify the corresponding probability μ_i. By Lemma 45, there exists (x_i, m_i, q_i) satisfying $2\sum x_i(m_i q_i)^{2j} = t_{2j}$ and $2\sum x_i m_i^2 q_i = 1$. For each triple add the triple $(m_i, q_i, \beta_i = 1)$ to the distribution. Let B be the set of indices i such that $\beta_i = 1$ and B^c be the set of indices i such that $\beta_i = 0$. Let $z = \sum_{i\in B} x_i\gamma + \sum_{i\in B^c} x_i(1-\gamma)$. We now define μ by assigning probabilities to triples (m_i, q_i, β_i). If $i \in B^c$, let $\mu_i = x_i\gamma/z$. If $i \in B$, let $\mu_i = x_i(1-\gamma)/z$. Note $\sum \mu_i = 1$, and therefore $(\mu, 0)$ is a well-defined ROC family with

$$x = 1/\left(\sum_{i\in B^c} \mu_i m_i^2 q_i + 2\sum_{i\in B} \mu_i m_i^2 q_i\right) = z.$$

Note $c(j) = c_j/2$ by construction. For $G \sim ROC(n, d, \mu, 0)$ and $d = o\left(n^{\frac{1}{k-1}}\right)$, Corollary 33 implies that $\mathsf{E}[C_j(G)] = \frac{c_j}{2} nd + o(nd)$. The statement follows.

For the other direction, there is some ROC family that achieves the limit. Since the number of j-cycles grows with nd for all j, we must have $a = 0$. Let $(\mu, 0)$ be the ROC family. Let $\gamma = x\sum_{i\in B^c} \mu_i m_i^2 q_i$, and so $1-\gamma = 2x\sum_{i\in B} \mu_i m_i^2 q_i$. For each $i \in B^c$, let $x_i = x\mu_i/\gamma$. Note $\sum_{i\in B^c} x_i m_i^2 q_i = 1$, and so by Lemma 44, the vector with $s_j = \sum_{i\in B^c} \mu_i(m_i q_i)^j$ satisfies the Stietljes condition. For each $i \in B$, let $x_i = x\mu_i/(1-\gamma)$. Note $2\sum_{i\in B} x_i m_i^2 q_i = 1$, and so by Lemma 45, the vector with $t_{2j} = \sum_{i\in B} \mu_i(m_i q_i)^{2j}$ satisfies the Stietljes condition. Theorem 27 implies that $c_j/2$ is the appropriate combination of s_j and t_j. □

4.3.3 Simple criterion for the the Stieltjes condition.

The following lemma provides a convenient criterion that implies the Stieltjes condition. In particular, we use this show that the limit of hypercube sequence is totally k-achievable (Theorem 2).

Lemma 46. *Let $s_1, s_2, \ldots s_k$ be a vector with $s_1 > 0$ satisfying $s_x s_y < s_a s_b$ for all $1 \le a < x \le y < b$. Then there exists $s_0 > 0$ such that $(s_0, s_1, s_2, \ldots s_k)$ satisfies the Stieltjes condition.*

The following is the key lemma for proving Lemma 46.

Lemma 47. *Let $s_1, s_2, \dots s_k$ be a vector with $s_1 > 0$ and $s_x s_y < s_a s_b$ for all $1 \le a < x \le y < b$. Let H be the $\lfloor \frac{k+1}{2} \rfloor \times \lfloor \frac{k+1}{2} \rfloor$ with $H_{ij} = s_{i+j-1}$. Then all leading principal minors of H have positive determinant.*

Proof. Let H_k denote the k^{th} leading principal minor of H. We show that $det(H_k) > 0$ by induction on k. Note $det(H_1) = s_1 > 0$. Next assume $det(H_{k-1}) > 0$. Write $H_k = AB$ where A and B are in the form displayed here.

$$H_k = \begin{pmatrix} s_1 & s_2 & s_3 & \dots & s_k \\ s_2 & s_3 & & & \vdots \\ s_3 & & & & \vdots \\ \vdots & & & \ddots & \vdots \\ s_k & \dots & \dots & \dots & s_{2k-1} \end{pmatrix} = \begin{pmatrix} s_1 & s_2 & \dots & s_{k-1} & 0 \\ s_2 & & & \vdots & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ s_{k-1} & \dots & \dots & s_{2k-3} & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & \dots & 0 & x_1 \\ 0 & 1 & & \vdots & x_2 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & \dots & \dots & 1 & x_{k-1} \\ s_k & s_{k+1} & \dots & s_{2k-2} & s_{2k-1} \end{pmatrix}.$$

Note $det(A) = det(H_{k-1})$, which is positive by the inductive hypothesis. It follows there exists a unique solution of real values $x_1, x_2, \dots x_{k-1}$ so that $H_k = AB$. Since $det(H_k) = det(A)det(B)$ and $det(A) > 0$, it suffices to show that $det(B) > 0$ to prove the inductive hypothesis.

Note $det(B) = s_{2k-1} - L$ where

$$L = \begin{pmatrix} s_k & s_{k+1} & \dots & s_{2k-2} \end{pmatrix} \begin{pmatrix} x_1 & x_2 & \dots & x_{k-1} \end{pmatrix}^T.$$

By construction of A and B,

$$\begin{pmatrix} s_1 & s_2 & \dots & s_{k-1} \\ s_2 & & & \vdots \\ \vdots & & \ddots & \vdots \\ s_{k-1} & \dots & \dots & s_{2k-3} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{k-1} \end{pmatrix} = \begin{pmatrix} s_k \\ s_{k+1} \\ \vdots \\ s_{2k-2} \end{pmatrix}. \tag{21}$$

For $i \in [0, k-2]$, define

$$\alpha_i = \frac{s_{k+i}}{s_{i+1} + \dots + s_{k-1+i}},$$

and let $\alpha = \max_{i \in [0,k-2]} \alpha_i$. Therefore for all $i \in [0, k-2]$

$$x_{i+1} s_{k+i} = \alpha_i x_{i+1}(s_{i+1} + \dots + s_{k-1+i}) \le \alpha x_{i+1}(s_{i+1} + \dots + s_{k-1+i}).$$

Summing the above equation over all $i \in [0, k-2]$ and applying equality (21) yields

$$L \le \alpha(s_k + s_{k+1} + \dots + s_{2k-2}).$$

To prove $det(B) = s_{2k-1} - L > 0$ we show that for all $i \in [0, k-2]$, $\alpha_i(s_k + s_{k+1} + \dots + s_{2k-2}) < s_{2k-1}$, or equivalently

$$s_{k+i}(s_k + s_{k+1} + \dots + s_{2k-2}) < s_{2k-1}(s_{i+1} + \dots + s_{k-1+i}). \tag{22}$$

Note by assumption $s_{k+i} s_{k+j} < s_{2k-1} s_{i+j+1}$ for all $i, j \in [0, k-2]$. Therefore the j^{th} term on the left side of (22) is less than the j^{th} term on the right side of (22), and so (22) holds. ∎

Proof. (of Lemma 46.) Let $H^{(0)}$ and $H^{(1)}$ be defined for the vector $(s_0, s_1, \dots s_k)$ as in Definition 6. Lemma 47 implies that all leading principal minors of $H^{(1)}$ have positive determinant. It remains to show that there exists $s_0 > 0$ such that all leading principal minors of $H^{(0)}$ have positive determinant.

Let H' be $H^{(0)}$ with the first row and column deleted. The i^{th} leading principal determinant of $H^{(1)}$ has the form $s_0 h_i + b_i$ where h_i is the $(i-1)^{st}$ principal determinant of H'. (Taking the determinant via expansion of the first row makes this clear.) Note that Lemma 47 applied to the vector $(s_2, s_3, \dots s_k)$ guarentees that each $h_i > 0$. Therefore, it is possible to pick s_0 sufficiently large such that all principal determinants $s_0 h_i + b_i$ of $H^{(1)}$ are positive. □

5 ROC achievable limits

5.1 Hypercube

In this section, we prove Theorem 2, which states that the limit of the hypercube sequence is totally k-achievable, and discuss the limits of several related graph sequences. First we provide the ROC parameters which achieve the small k-limits of the hypercube.

Remark 48. *The ROC family $(\mu, 1/2)$ where μ is the distribution with support size one on $(8, 1/4, 1)$ achieves the 6-limit of the hypercube sequence. To achieve longer limits, the distribution will have larger support.*

We now prove Theorem 2. Recall from Lemma 20 that the sparsity exponent of the hypercube sequence is 1/2. Therefore, to prove the theorem we apply Theorem 5, which states the vector $(w_3, w_4, \dots w_k)$ can be achieved by ROC if the normalized cycle count vector $(c_3, c_4, \dots c_k)$ corresponding to the transform T can be extended to satisfy the Stieltjes condition. The following lemma gives the cycle vector for the hypercube.

Lemma 49. *Recall from Lemma 20 that the limit of the hypercube sequence (G_d) is $(w_3, w_4, \dots)$ where*

$$w_j = \begin{cases} (j-1)!! & \text{for } j \text{ even} \\ 0 & \text{for } j \text{ odd.} \end{cases}$$

For T the cycle transform given in Definition 8, $T((0, s_2, 0, s_3, 0, \dots)) = (0, w_4, 0, w_6, 0, \dots)$ where $s_1 = 1$ and $s_n = (n-1)\sum_{j=1}^{n-1} s_j s_{n-j}$.

Remark 50. *This sequence is A000699 in OEIS:* $1, 1, 4, 27, 248, 2830, \dots$

Proof. (of Lemma 49) Note that the hypercube is vertex transitive so the sequence of hypercubes is essentially k-locally regular. Therefore, Theorem 31 implies $C((0, w_4, 0, w_6, 0, \dots)) = (c_3, c_4, c_5, \dots)$ where

$$c_i = \lim_{d\to\infty} \frac{C_i(G_d)}{2^d d^{i/2}}$$

where $C_i(G_d)$ is the normalized number of i cycles at a vertex. Instead of applying polynomial operations to the vector $(0, w_4, 0, w_6, 0, \dots)$ to obtain $(c_3, c_4, c_5, \dots)$ we directly compute $C_k(G_d)$, the number of cycles in the d-dimensional hypercube graph. As in Lemma 20, we think of k-walks on the hypercube as length k strings where the i^{th} character indicates which of the d coordinates is changed on the i^{th} edge of the walk. For closed walks each coordinate that is changed must be changed back, so each coordinate that appears in the string must appear an even number of times. For $1 \le i \le k/2$, let Z_i be the number of such strings of length k that involve i coordinates and correspond to a k-cycle on the hypercube graph. Since there are d coordinates $Z_i = \Theta(d^i) = o(d^{k/2})$ for $i < k/2$ and so

$$C_k(G_d) = nZ_{k/2} + o\left(nd^{k/2}\right)$$

where $n = 2^d$ is the number of vertices.

We compute $Z_{k/2}$ by constructing a correspondence between length k strings with $k/2$ characters each appearing twice that represent cycles and irreducible link diagrams. A link diagram is defined as $2n$ points in a line with n arcs such that each arc connects precisely two distinct points and each point is in precisely one arc. The arcs define a complete pairing of the interval $[1, 2n]$. A link diagram is reducible if there is a subset of $j < n$ arcs that form a complete pairing of a subinterval of $[1, 2n]$ and irreducible otherwise. Let S be the set of k length strings in which the characters $1, 2, \dots k/2$ each appear twice and the first appearance of character i occurs before the first appearance of j for all $i < j$. Let L be the set link diagrams L on k points. Let $\overline{S} \subseteq S$ be the subset of strings that correspond to cycles on the hypercube and let $\overline{L} \subseteq L$ be the set of irreducible link diagrams.

We construct a bijection $f : S \to L$ and show f restricted to $\overline{S}$ gives a bijection between $\overline{S}$ and $\overline{L}$. Given $s \in S$, we construct a corresponding link diagram $f(s) \in L$ by labeling k points so that the i^{th} point is labeled with the i^{th} character of s and then drawing an arc between each pair of points with the same label. Note f is a bijection. (To produce $f^{-1}(\ell)$ label the arcs $1, 2, \dots k/2$ by order of their left endpoints. Label each point with the label of its arc and read off the string of the labels.) It remains to show that $f(s) \in \overline{L}$ if and only if $s \in \overline{S}$. We prove the contrapositive. Suppose $s \notin \overline{S}$. Then the hypercube closed walk corresponding to s is not a cycle. Therefore there exists j and steps $i, i+1, \dots i+j$ of the walk that make a $j/2$ cycle. (Here by convention traversing an edge twice is a 2-cycle.) Since $i, i+1, \dots i+j$ form a cycle, each coordinate that was changed between step i and step $i+j$ must have been changed back. Therefore each character that appears in the interval $[i, i+j]$ appears appears twice. It follows $f(s)$ has a complete pairing of the subinterval $[i, i+j]$ and therefore is reducible. Next suppose $\ell \notin \overline{L}$. Then there exists a subinterval $[i, i+j] \neq [1, 2n]$ with a complete pairing. Therefore the walk corresponding to $f^{-1}(\ell)$ is not a cycle because the walk visits the same vertex before step i and after step $i+j$. It follows that $f^{-1}(\ell) \notin \overline{S}$.

Thus $|\overline{L}| = |\overline{S}|$. See [37] for a proof that $|L| = s_{k/2}$. There are $d(d-1)\dots(d-k/2+1) = d^{k/2} + o\left(d^{k/2}\right)$ ways to select the $k/2$ coordinates in the order they will be changed. Thus $Z_{k/2} = d^{k/2}|S| + o\left(d^{k/2}\right) = s_{k/2}d^{k/2} + o\left(d^{k/2}\right)$, and therefore

$$C_k(G_d) = s_{k/2}nd^{k/2} + o\left(nd^{k/2}\right).$$

□

Lemma 51 (from [37]). *Let $s_1 = 1$ and $s_n = (n-1)\sum_{j=1}^{n-1} s_j s_{n-j}$. Let $s_1 = 1$ and $s_n = (n-1)\sum_{j=1}^{n-1} s_j s_{n-j}$. Then*

$$s_{n+1} > (2n+1)s_n \quad \textit{for} \quad n \ge 4$$

$$s_{n+1} < (2n+2)s_n \quad \textit{for} \quad n \ge 1.$$

Lemma 52. *Let $s_1 = 1$ and $s_n = (n-1)\sum_{j=1}^{n-1} s_j s_{n-j}$. Let $s_1 = 1$ and $s_n = (n-1)\sum_{j=1}^{n-1} s_j s_{n-j}$. Then there exists $s_0 > 0$ such that $(s_0, s_1, s_2, \dots s_k)$ satisfies the Stieltjes condition.*

Proof. We apply Lemma 46 which says a vector with $s_x s_y < s_a s_b$ for all $1 \le a < x \le y < b$ can be can be extended to satisfy the Stieltjes condition. We show this conditions holds for the infinite vector.

First we consider the case when $y \ge 4$. By Lemma 51

$$s_b > \frac{(2b-1)!!}{(2y-1)!!} s_y \quad \text{and} \quad s_x < \frac{(2x)!!}{(2a)!!} s_a.$$

Note $\frac{(2x)!!}{(2a)!!} < \frac{(2b-1)!!}{(2y-1)!!}$, and therefore

$$s_x s_y < \frac{(2x)!!}{(2a)!!} s_a s_y < \frac{(2b-1)!!}{(2y-1)!!} s_a s_y < s_a s_b.$$

We consider the remaining three cases separately. For $a = 1, x = 2, y = 3$, $b \ge 4$ so $s_b \ge 27$. Therefore $s_x s_y = 4 < 27 \le s_a s_b$. For $a = 1, x = 3, y = 3$, $b \ge 4$ so $s_b \ge 27$. Therefore $s_x s_y = 16 < 27 \le s_a s_b$. For $a = 1, x = 2, y = 2$, $b \ge 3$ so $s_b \ge 4$. Therefore $s_x s_y = 1 < 4 \le s_a s_b$. □

Proof. (of Theorem 2) Follows directly from Lemma 49, Lemma 52, and Theorem 5. □

Remark 53. *The limit of the sequence of hypercubes is not fully achievable.*

Proof. For a ROC family $(\mu, 1/2)$ with $m = \max_i m_i$, the w_k coordinate in the limit vector is at most $x(2m)^k$. However the w_k coordinate in the hypercube sequence is $(k-1)!! = \Theta\left(\left(\frac{k}{e}\right)^{k/2}\right)$. Therefore there is no μ which achieves the full hypercube limit vector. □

5.1.1 Generalized hypercubes

Two generalizations of the hypercube have the same limit and therefore are also totally k-achievable.

Corollary 54 (Hypercube generalizations)**.** *The following sequences of graphs (G_d) converge with sparsity exponent 1/2 to the same limit as the hypercube sequence.*

1. *(Hamming generalization) Let G_d be the graph on vertex set $\{0, 1, \dots, k-1\}^d$ where two vertices are adjacent if the Hamming distance between their labels is one.*

2. *(Cayley generalization) Let G_d be the graph on vertex set $\{0, 1, \dots, k-1\}^d$ where two vertices are adjacent if their labels differ by a standard basis vector.*

Proof. Since the Hamming and Cayley sequences are locally regular, it suffices to show that the sequences have sparsity exponent 1/2 and the same vector of normalized cycle counts as the hypercube. Let D denote the degree of the graph G_d, so for the Hamming graph $D = d(k-1)$ and for the Cayely graph $D = 2d$.

First we show that both sequences have sparsity exponent 1/2 by showing that each vertex is in $O\left(D^{i/2}\right)$ i-cycles (locally regularity guarantees the walk counts are of the same order). We count i-cycles by grouping them according to the number of coordinate positions changed during the cycle, as in the proof of Lemma 49. The highest order term comes from i cycles in which $i/2$ coordinates are changed. Therefore the number of i-cycles at each vertex is $O\left(d^{i/2}\right) = O\left(D^{i/2}\right)$ and it follows that the sparsity exponent is $1/2$.

Next we compute the cycle vector $c_i = \lim_{d\to\infty} \frac{C(G_d)}{nD^{i/2}}$. The number of i-cycles at a vertex that involve changing fewer than $i/2$ coordinates is $o(D^{i/2})$, so such cycles do not contribute to c_i. Therefore, while there are odd cycles in the Hamming and Cayley graphs, $c_i = 0$ for i odd. We now count the number of i-cycles at a vertex that involve changing $i/2$ coordinates. As described in Lemma 49 there are $s_{i/2}d^{i/2}$ ways to select $i/2$ coordinates and change them in a manner that corresponds to a cycle. In the hypercube, there is only one way to change a single coordinate, so the total number of cycles at a vertex is $s_{i/2}d^{i/2}$.

In the Hamming graph there are $k-1$ ways to change a coordinate since there are k possibilities for each coordinate. Therefore, for the Hamming sequence and i even

$$c_i = \lim_{d\to\infty} \frac{s_{i/2}(k-1)^{i/2}}{nD^{i/2}} = \lim_{d\to\infty} = s_{i/2}.$$

In the Cayley graph there are two ways to change a single coordinate (either add one or subtract one). Therefore, for the Cayley sequence and i even

$$c_i = \lim_{d\to\infty} \frac{s_{i/2}2^{i/2}}{nD^{i/2}} = \lim_{d\to\infty} = s_{i/2}.$$

□

Remark 55. *The above corollary shows that same ROC family (μ, a) achieves the k-limit of the sequence of hypercubes, and the closely related Hamming and Cayley generalizations. While these sequences all have the same ROC family limit object, the ROC family can produce sequences of ROC graphs (G_d), $G_d \sim ROC(n_d, d_d, \mu, a)$, unique to each of these settings by varying relationship between n_d and d_d. A sequence with $n_d = 2^d$ and $d_d = d$ will match the edge density and unnormalized walk counts of the hypercube, whereas a sequence with $n_d = k^d$ and $d_d = d(k-1)$ or $d_d = 2d$ will match the edge density and unnormalized walk counts of the Hamming or Cayley generalization respectively.*

5.2 Rook graph

We now prove Theorem 3 which states that the limit of the sequence of rook's graphs is fully achievable.

Proof. (of Theorem 3.) Recall from Lemma 21 that the sequence of (G_k) has sparsity exponent 1 and converges to the vector with $w_i = 2^{2-i}$. By Theorem 27, the ROC family with $a = 1$ and μ the distribution that selects $m = 1/2$ and $q = 1$ with probability 1 achieves this limit. □

5.3 Erdős-Rényi sequences

We consider ROC approximations of the sequences of Erdős-Rényi graphs given in Lemma 24.

Theorem 56. *Let $\ell > 1$. Let $(G_n) \sim G(n^{2\ell}, n^{2\ell-2})$.*

1. *For $k < 2\ell$, the k-limit of (G_n) is achieved by any ROC family with $a < 1/2$.*

2. *For $k \geq 2\ell$, the k-limit of (G_n) is not k-achievable by any ROC family. However, for any $\varepsilon > 0$, there exists a ROC k-achievable vector that is L_∞ distance at most ε from the k-limit.*

3. *The sparsity exponent of (G_n) is 1 and the limit is $(0, 0, \dots)$. This limit is not ROC fully achievable. However, for any $\varepsilon > 0$, there exists a ROC fully achievable vector that is L^∞ distance at most ε from $(0, 0, \dots)$.*

Proof. For $k < 2\ell$ the sparsity exponent of (G_n) is $1/2$ and the k-limit is $(w_3, w_4, \dots w_k)$ where $w_i = 0$ for i odd and $w_i = Cat_{i/2}$ for i even. By Theorem 27, this is the limit for any ROC family with $a < 1/2$.

For $k = 2\ell$, the k-sparsity exponent of (G_n) is $1/2$ and the k-limit is $(w_3, w_4, \dots, w_{k-1}, \overline{w}_k)$ where $w_i = 0$ for odd i, $w_i = Cat_{i/2}$ for even i, and $\overline{w}_k = w_k + 1$. By Theorem 27, in order to approximate the vector it is necessary to have μ be such that $x \sum \mu_i (m_i q_i)^j = c_j$ where $(c_3, c_4, \dots c_k) = T(w_3, w_4, \dots, w_{k-1}, \overline{w}_k)$ is the cycle transform, so $c_j = 0$ for $j < k$ and $c_k = 1$. Since $\mu_i, q_i, m_i > 0$, $c_k = 1$ implies $c_j \neq 0$ for all $j < k$. Therefore, the vector cannot be achieved exactly by ROC.

We now show that it is possible to achieve a vector that is arbitrarily close to the desired vector with respect to the L_∞ metric. Note that for μ the distribution on one point $m = \delta^{\frac{1-k}{k-2}}$ and $q = \delta$, the resulting ROC family $(\mu, 1/2)$ has

$$c_j = m^{j-2} q^{j-1} = \delta^{\frac{(j-2)(1-k)}{k-2} + j - 1}.$$

Therefore $c_k = 1$ and c_j for $j < k$ can be made arbitrarily small by decreasing δ. To achieve L_∞ distance ε, choose δ small enough so that $\max_{j<k} w_j = max_j T(c_3, c_4, \dots c_k)_j < \varepsilon$.

For $k > 2\ell$, the sparsity exponent of (G_n) is $\frac{k-\ell-1}{k-2}$ and the k-limit is $(w_3, w_4, \dots, w_k)$ where $w_i = 0$ for $i < k$, $w_k = 1$. Therefore, by Theorem 27, to approximate the vector we likewise need μ be such that $x \sum \mu_i (m_i q_i)^j = w_j$ where $c_j = w_j = 0$ for $j < k$ and $w_k = c_k = 1$, and the result is as in the previous case.

Similarly, we can approximate the vector $(0, 0, \dots)$ with sparsity exponent 1 up to arbitrarily small error with respect to the L_∞ distance. Note that for μ the distribution on one point m and q, the ROC family $(\mu, 1)$ has $w_k = m^{k-2} q^{k-1}$. Therefore it is possible to achieve error ε by selecting m and q such that $\max_k m^{k-2} q^{k-1} < \varepsilon$. □

6 Discussion

6.1 Limitations of the ROC model

We have shown that the ROC model is an insightful limit object for many sequences of graphs; the model is succinct and can be easily sampled to produce graphs with the same normalized walk counts as the sequence up to terms that disappear as the size of the sampled graph grows. Theorems 4, 5, 40 and 41 give necessary and sufficient conditions for when a limit sequence can be achieved. The natural next question is whether all graphs sequences converge to a limit that can be achieved by a ROC family. The answer to this question is no. There are both sequences of graphs that are not convergent in any of the senses we have defined, and convergent sequences of graphs with limits that are not achievable by a ROC family. We discuss such examples in this section.

Non-convergent graph sequences. Not all sequences of graphs converge or have a convergent subsequence. Consider the following sequence of ROC graphs drawn from different ROC families.

Example 57. *Let $a \in [1/2, 1]$ and let μ_i be the distribution on one point $m_i = i$ and $q_i = 1$. Let (G_i) be a sequence of graphs with $G_i \sim ROC(n_i, d_i, \mu_i, a)$ such that d_i satisfies the degree conditions given in Definition 5. The sequence (G_i) is not k-convergent for any k and is not fully convergent.*

Proof. By Theorem 27, $\mathsf{E}[W_3(G_i)] = m_i q_i^2 n_i d_i^{1+a(k-2)} + o\left(n_i d_i^{1+a(k-2)}\right)$. Therefore $W_3(G_i, a) = i + \varepsilon(i)$ and for all $\alpha > a$ $W_3(G_i, \alpha) = 0 + \varepsilon(i)$ where $\varepsilon(i)$ is an error term that vanishes with high probability as i tends to infinity. It follows that almost surely the sparsity exponent and k-sparsity exponent are a and the sequence $W_3(G, a)$ does not converge. Thus, the sequence is almost surely not k-convergent or fully convergent. □

Limit sequences that are not achievable by any ROC model We give a sequence of graphs with increasing degree that converges to a limit that is not achievable by any ROC family and provide a method for producing such sequences. First we give a necessary condition on (w_3, w_4, w_5, w_6) for it to appear as the prefix of a limit vector achievable by a ROC family.

Lemma 58. *If (w_3, w_4, w_5, w_6) is a prefix of a k-limit that can be achieved by ROC family with sparsity exponent $> 1/2$, then*

$$w_3^2 w_6 \geq w_5^3.$$

Proof. By Theorem 4, there exists $\gamma \in [0, 1]$, s_0, s_1, and $s_2 \leq 1$ such that $(s_0, s_1, \ldots s_k)$ satisfies the Stieltjes condition and $w_j = s_j \gamma$ for j odd. This condition implies that $H_{2s}^{(0)}$ and $H_{2s+1}^{(1)}$ as defined in Definition 6 are positive semidefinite. It follows that the principal minors

$$\begin{pmatrix} s_4 & s_5 \\ s_5 & s_6 \end{pmatrix} = \begin{pmatrix} s_4 & \frac{w_5}{\gamma} \\ \frac{w_5}{\gamma} & s_6 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} s_3 & s_4 \\ s_4 & s_5 \end{pmatrix} = \begin{pmatrix} \frac{w_3}{\gamma} & s_4 \\ s_4 & \frac{w_5}{\gamma} \end{pmatrix}$$

of $H_{2s}^{(0)}$ and $H_{2s+1}^{(1)}$ respectively have non-negative determinant. Therefore

$$\gamma^2 s_4 s_6 \geq w_5^2 \quad \text{and} \quad \gamma^2 s_4^2 \leq w_3 w_5,$$

and so $\gamma s_6 \geq \frac{w_5^{3/2}}{\sqrt{w_3}}$. Since $w_6 \geq \gamma s_6$, the statement follows. □

Next we show how to construct a sequence of graphs with increasing degree that fails this condition. This construction is due to Shyamal Patel.

Lemma 59. *Let G_0 be a graph. We construct a sequence G_i as follows. Let G_i be the graph with adjacency matrix $\begin{pmatrix} A_{i-1} & A_{i-1} \\ A_{i-1} & A_{i-1} \end{pmatrix}$ where A_{i-1} is the adjacency matrix of G_{i-1}. Let $w_j = W_j(G_0, 1)$. Then for each i,*

$$W_j(G_i, 1) = w_j,$$

and so (G_i) converges to $(w_3, w_4, w_5, \ldots)$ with sparsity exponent 1.

Proof. Let G_0 be a graph on n vertices with average degree d. Let A_0 be the adjacency matrix of G_0 and let $\lambda_1 \geq \lambda_2 \geq \dots \lambda_\ell$ be the non-zero eigenvalues of A_0. Note that if λ is an eigenvalue of A_{i-1} with eigenvector v, then 2λ is an eigenvalue of A_i with eigenvector $\begin{bmatrix} v \\ v \end{bmatrix}$. Since the adjacency matrix A_i has the same rank as the adjacency matrix of A_{i-1}, the set of non-zero eigenvalues of A_i is precisely the set of non-zero eigenvalues of A_{i-1} in which each is doubled. Therefore A_i has non-zero eigenvalues $2^i\lambda_1, 2^i\lambda_2, \dots 2^i\lambda_\ell$. Note G_i has $G_0 2^i$ vertices and average degree $2^i d$. Therefore

$$W_j(G_i, 1) = \frac{\sum_{b=1}^{\ell}(2^i\lambda_b)^j}{2^i n(2^i d)^{j-1}} = \frac{\sum_{b=1}^{\ell}(\lambda_b)^j}{nd^{j-1}} = w_j.$$

□

The lemma implies that if there is a graph with $W_j(G, 1) = w_j$, then there is a sequence of graphs (G_i) with increasing degree that converges to this limit with sparsity exponent 1. Taking G_0 to be a girth four graph yields a sequence (G_i) with a limit vector that violates the condition of Lemma 58. This sequence is dense since $d_i = \Theta(n_i)$. However, we can construct a sparser sequence (G_i') from (G_i) with the same limit by taking each G_i' to be the union of disjoint copies of G_i.

Lemma 60. *Let (G_i) be a convergent sequence of graphs with sparsity exponent α. Let (t_i) be a sequence of positive integers, and let (G_i') be a graph sequence in which G_i' consists of t_i disjoint copies of G_i. Then (G_i') achieves the same limit as G_i with the same sparsity exponent.*

Proof. Note that G_i and G_i' both have average degree d_i and the number of vertices in G_i' is $n_i' = t_i n_i$. Note also that $W_k(G_i') = t_i W_k(G_i)$. It follows that

$$W_k(G_i', \alpha) = \frac{W_j(G_i')}{n_i' d_i'^{1+a(j-2)}} = \frac{W_j(G_i)}{n_i d_i^{1+a(j-2)}}.$$

□

We use Lemmas 58 and 59 to construct a family of sequences with arbitrary sparsity that are not achievable by the ROC model. This example implies that there is no class of densities for which the ROC model can capture all 6-limits of sequences with the specified density.

Example 61. *Let G_0 be the five cycle. Then the sequence (G_i) defined as in Lemma 59 converges to a limit that cannot be achieved by any ROC family. This limit $(0, 3/4, 1/8, 5/8)$ is at constant distance from any achievable limit. Moreover, there exists a sequence (G_i') with the same limit and $d_i' = f(n_i')$ for any function $f(n) = o(n)$. To see this, apply Lemma 60 to the sequence (G_i) with $t_i = f^{-1}(n_i)/n_i$.*

6.2 ROC as a model for real-world graphs.

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 (e.g. school, work, or a particular hobby); the conceptualization of social networks as overlapping communities has been studied in [30], [42]. 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], [20], [6].

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.

Moreover, the ROC model could be used as a null hypothesis for testing properties of a real-world networks known to have community structure. It is established practice to compare real-world graphs to various random graph models to understand the non-random aspects of its structure ([12, 35, 26, 28]). The ROC model is particularly well-suited to be the null hypothesis graph for graphs with known community structure. Comparing such a network to a ROC network would differentiate between properties of the network that are artifacts of community structure and those that are unique to the graph.

6.3 ROC as a limit object

We have seen that the ROC model provides a sampleable approximation of the limits of many sparse graph sequences, in particular the hypercube sequence. Our metric was defined in terms of a vector of closed walk counts of each length appropriately normalized. This vector is a natural choice because closed walk counts are equivalent to the moments of the eigenspectrum, and the normalization factor encodes average density of local neighborhoods. Our findings suggest that measuring closed walk counts and approximating with the ROC model is a promising beginning to a complete theory for describing the limits of sparse graph sequences (in particular those with roughly uniform degree and are not captured by graphexes). We end with a discussion of future directions that illustrate the potential of the ROC model and address current limitations of the theory.

Distance and convergence. Our notion of convergence based on normalized closed walk count vectors differs from other notions of graph convergence in two key ways. First, our theory does not provide an inherent metric for describing the distance between two graphs. The normalization factor used to determine the convergence of a sequence of graphs depends on the rate of growth of the closed walk counts in the sequence. Therefore, it not clear which normalization factor α to use when comparing the closed walk vectors of just two graphs. Second, due to this flexibility in normalization parameter α, the space of all vectors of normalized closed walk counts is unbounded, and so it is possible to construct sequences of graphs with no convergent subsequence (as in Section 6.1). In contrast the set of local profiles and the set of graphons are compact, so every sequence of graphs in these settings has a convergent subsequence. Further investigation is necessary to meaningfully extend the ROC theory to the context of approximating a small set of graphs rather than a sequence, and to the context of non-convergent sequences.

Capturing cuts. While a graph H drawn from the ROC model may capture the closed walk counts of a graph G, there is no guarantee that H and G will have similar cuts. (The in the local profile approach for bounded degree graphs also succeeds at encoding local properties and fails to capture the global property of cuts.) For example, consider a convergent sequence of connected graphs (G_i) and a sequence of graphs (G'_i) where G'_i is a collection of disjoint copies of G_i. Lemma 60 implies (G_i) and (G'_i) have the same limit; however the cuts in these sequences greatly differ. Moreover the cuts of a ROC graph drawn from the family that achieves the limit need not have cuts that match either G_i or G'_i.

In general, even if the moments of the eigenspectra of two graphs match, their spectral gaps and precise set of eigenvalues may greatly differ. In Appendix C we discuss a different approximation of the spectrum of the hypercube graph. It is not of constant size (the size of the approximation grows with d for a hypercube of size 2^d), but it captures the d distinct eigenvalues of the hypercube precisely (and therefore the minimum cut). On the other hand, the approximation does not preserve information about the multiplicities of the eigenvalues, and hence does not capture the moments.

An extension of the ROC model. We imagine the following extension to the ROC model that has the potential to encode information about the cuts of a graphs and give a finer grained approximation of local structure while also maintaining the approximation of closed walk counts. Begin with a partition of the vertex set, and for each community type specify a distribution over partition classes. Then, when adding a community to a ROC graph, instead of selecting community membership from the entire vertex set with equal probability, select vertices for the community based on the corresponding distribution over partition classes. This modification has the potential to better approximate cuts because it is possible to control the number of edges between partition classes.

Moreover, the above modification will likely be a better approximation for graphs that are not necessarily close to locally regular. Currently a ROC approximation produces a graph in which each vertex is in approximately the number of closed walks as an average vertex in the target graph. However, the target graph could be made up of several types of vertices where all vertices of a given type have the same local closed walk count vector. A ROC approximation captures the weighted average of these vectors, but does not retain information about the distribution over such vectors. An expanded theory, perhaps including the above modification, could create graph approximations that capture the distribution of local closed walk counts vectors at each vertex.

Achievability of all limits. As demonstrated in Section 6.1, not all limit sequences can be achieved by a ROC model. In particular, the model may not be able to capture the limits of sequences of girth five graphs because the density of the communities need to produce many five cycles will also produce many three and four cycles. This problem could be resolved by generalizing the model so that communities may have structure other than E-R random graphs. Alternately, the aforementioned approach of adding ROCs between partition classes might provide sufficient flexibility to achieve a wider range of limits.

6.4 Additional open questions.

1. 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)?

2. Can the DROC model be extended to produce graphs with arbitrary clustering coefficients and degree distributions (that have long upper tails)? A modification of the DROC model could be that vertices with higher target degrees are more likely to join each community.

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 [38, 20]. 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. A ROC graph H that approximates a target graph G has similar closed walk counts as G. To what extent does this similarly imply that algorithms will behave similarly on G and H? For instance, can we analyze the behavior of random walks or percolation of ROC graphs? How does this behavior compare to the behavior of the same process on other graphs with the same closed walk counts?

5. Moreover, the asymptotic thresholds for properties of ROC graphs have yet to be studied. (See [17] for a survey on E-R random graphs.) Which phase transitions appearing in E-R random graphs also appear in ROC graphs? Does every nontrivial monotone property have a threshold?

Acknowledgements

The authors would like to thank Shyamal Patel for his construction of the class of graphs with non-achieveable limit given in Lemma 59 and Winston Li for help with real-world networks. Lutz Warnke for helpful conversations and pointers to relevant work, and Victor Veitch for his correspondence which clarified our understanding of graphexes. Finally, we thank László Lovász for inspiring us to think about the sequence of hypercube graphs.

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] Itai Benjamini and Oded Schramm. Recurrence of distributional limits of finite planar graphs. *Electron. J. Probab.*, 6:13 pp., 2001.

[8] Christian Borgs, Jennifer T Chayes, Henry Cohn, and Victor Veitch. Sampling perspectives on sparse exchangeable graphs. *arXiv preprint arXiv:1708.03237*, 2017.

[9] Christian Borgs, Jennifer T Chayes, Henry Cohn, and Yufei Zhao. An l^p theory of sparse graph convergence i: limits, sparse random graph models, and power law distributions. *arXiv preprint arXiv:1401.2906*, 2014.

[10] Andries E Brouwer and Willem H Haemers. Strongly regular graphs. In *Spectra of Graphs*, pages 115–149. Springer, 2012.

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

[12] Giovanni Ciriello and Concettina Guerra. A review on models and algorithms for motif discovery in protein–protein interaction networks. *Briefings in Functional Genomics and Proteomics*, 7(2):147–156, 2008.

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

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

[15] Péter E Frenkel. Convergence of graphs with intermediate density. *arXiv preprint arXiv:1602.05937*, 2016.

[16] Alan Frieze and Ravi Kannan. The regularity lemma and approximation schemes for dense problems. In *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, pages 12–20. IEEE, 1996.

[17] Alan Frieze and MichałKaroński. *Introduction to Random Graphs.* Cambridge University Press, 2015.

[18] Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. *Physical review E*, 65(2), 2002.

[19] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[20] 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.

[21] László Lovász and Santosh Vempala. Personal Communication, 2014.

[22] 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.

[23] 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.

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

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

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

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

[28] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2):026118, 2001.

[29] 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.

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

[31] 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.

[32] Josef Rukavicka. On generalized dyck paths. *The Electronic Journal of Combinatorics*, 18(1):P40, 2011.

[33] Comandur Seshadhri, Tamara G Kolda, and Ali Pinar. Community structure and scale-free collections of erdős-rényi graphs. *Physical Review E*, 85(5):056109, 2012.

[34] James Alexander Shohat and Jacob David Tamarkin. *The problem of moments*. Number 1. American Mathematical Soc., 1943.

[35] 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.

[36] Richard P Stanley. *Enumerative Combinatorics. Vol. 1, vol. 49 of Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1997.

[37] Paul R Stein. On a class of linked diagrams, i. enumeration. *Journal of Combinatorial Theory, Series A*, 24(3):357–366, 1978.

[38] 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.

[39] E Szemerédi. Regular partitions of graphs, colloques internationaux cnrs no 260–problemes combinatoires et théorie des graphes, orsay. *Computer Science Department Worcester Polytechnic Institute Worcester, MA 01609 USA*, pages 399–401, 1976.

[40] Victor Veitch and Daniel M Roy. Sampling and estimation for (sparse) exchangeable graphs. *arXiv preprint arXiv:1611.00843*, 2016.

[41] Erik Volz. Random networks with tunable degree distribution and clustering. *Physical Review E*, 70(5):056115, 2004.

[42] 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.

[43] 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 62. *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

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

□

Proof. (of Proposition 1.1) Let $\sigma_1 \dots \sigma_{rank(M)}$ denote the eigenvalues of M.

$$\begin{aligned}
\mathsf{E}[\#k\text{-cycles}] &= \sum_{i_1 \neq i_2 \dots \neq i_k} M_{i_1 i_2} M_{i_2 i_3} \dots M_{i_k i_1} \\
&\leq Tr(M^k) \\
&= \sum_{i=1}^{rank(M)} \sigma_i^k \\
&\leq rank(M) d^k.
\end{aligned}$$

□

B Connectivity of the ROC model

We describe the thresholds for connectivity for ROC graphs with one community type, $ROC(n, d, s, q)$. 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 asympotically almost surely. An event A_n happens a.a.s. if $\mathsf{Pr}[A_n] \to 1$ as $n \to \infty$.

Theorem 63. *For $(s-1)q(\ln n + c) \leq d \leq (s-1)qe^{sq}(1-\varepsilon)$, a graph from $ROC(n, d, s, q)$ a.a.s. has at most $(1+o(1))\frac{e^{-c}}{1-\varepsilon}$ isolated vertices.*

Proof. We begin by computing the probability a vertex is isolated,

$$\begin{aligned}
\mathsf{Pr}[v \text{ is isolated}] &= \sum_{i=0}^{\frac{nd}{s^2 q}} \mathsf{Pr}[v \text{ is in } i \text{ communities}](1-q)^{si} \\
&= (1+o(1)) \sum_{i=1}^{\frac{nd}{s^2 q}} \binom{\frac{nd}{s(s-1)q}}{i} \left(\frac{s}{n}\right)^i \left(1 - \frac{s}{n}\right)^{\frac{nd}{s(s-1)q} - i} e^{-sqi} \\
&\leq (1+o(1))\, e^{-\frac{d}{(s-1)q}} \sum_{i=0}^{\frac{nd}{s^2 q}} \left(\frac{de^{-sq+\frac{s}{n}}}{(s-1)q}\right)^i \\
&= (1+o(1))\, e^{-\frac{d}{(s-1)q}} \sum_{i=1}^{\frac{nd}{s^2 q}} \left(\frac{de^{-sq}}{(s-1)q}\right)^i \\
&= (1+o(1)) \left(e^{-\frac{d}{(s-1)q}}\right) \left(\frac{1}{1-\varepsilon}\right).
\end{aligned}$$

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

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

□

Theorem 64. *A graph from ROC(n, d, s, q) with $s = o(\sqrt{n})$ has no isolated communities a.a.s. if*

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

Proof. We construct a "community graph " and apply the classic result that $G(n, p)$ will a.a.s. have no isolated vertices when $p > (1+\epsilon)\log n/n$ for any $\epsilon > 0$[13]. 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 $\left(1 - \left(\frac{s}{n}\right)^2\right)^n$. Since communities are selected independently, the "community graph " is an instance of $G\left(\frac{nd}{s(s-1)q}, \left(1 - \left(\frac{s}{n}\right)^2\right)^n\right)$. By the classic result, approximating the parameters by $\frac{nd}{s^2q}, 1 - e^{s^2/n}$, 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

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

□

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 Approximating the hypercube's set of eigenvalues

The ROC model captures the first k moments of the eigenspectrum of the hypercube. To illustrate the difficultly of capturing the moments with a random model, here we an approximation method that preserves the set of eigenvalues of the hypercube (without multiplicity) and show this is not enough to also capture the moments. This approximation is a $2^d \times 2^d$ matrix M of rank d with entries in the interval $[0, 1]$ that has the same set of eigenvalues as the d-dimensional hypercube. A graph is produced by connecting vertices v_i and v_j with probability $M_{i,j}$.

The following well-known claim describes the eigenspectrum of the hypercube.

Proposition C.1. *Let A_d be the adjacency matrix of the d-dimensional hypercube graph. Then for $i \in \{0, 1, \ldots d\}$, $-d + 2i$ is an eigenvalue of A_d with multiplicity $\binom{d}{i}$.*

We approximate the adjacency matrix A_d of the d-dimensional hypercube graph by dividing the hypercube into layers. Layer i consists of the $\ell_i = \binom{d}{i}$ vertices whose labels have precisely i zeros. We let p_i be the fraction of edges between layer i and layer $i+1$. Each vertex in layer i has $d - i$ neighbors in layer $i + 1$, so $p_i = \frac{d-i}{\ell_{i+1}}$. We construct M_d, a block matrix in which the ij block has width ℓ_j and height ℓ_i, and each entry in the block matrix is the probability that two distinct randomly selected vertices from layer i and layer j are adjacent.

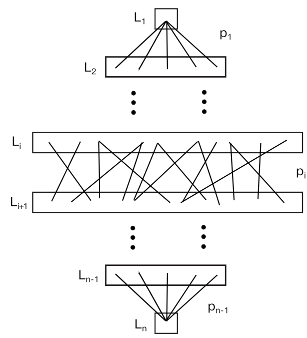

Figure 6: A graphical representation of a graph sampled from M_d.

Lemma 65. *The set of eigenvalues of M_d is precisely the set of eigenvalues of the d-dimensional hypercube.*

Proof. Let M_d^* be the $d \times d$ matrix obtained from M_d by the following procedure. For each of the $d+1$ blocks, replace the ℓ_i rows corresponding to block i with a single 1×2^d row equal to the sum of the ℓ_i rows. This yields a $d \times 2^d$ matrix. Delete all duplicate columns to obtain

$$M_d^* = \begin{bmatrix} 0 & 1 & & & \cdots & \\ n & 0 & 2 & & & \\ & n-1 & 0 & & & \\ & & & \ddots & & \\ \vdots & & & & 0 & n \\ & & & & 1 & 0. \end{bmatrix}$$

First we claim that if λ is an eigenvalue of M_d^*, then λ is also an eigenvalue of M_d. Suppose $v = (v_0, \ldots v_d)$ is an eigenvector of M_d^* with eigenvalue λ. Let v^* be the vector obtained by replacing each entry v_i with ℓ_i entries with value v_i/ℓ_i. Note v^* is an eigenvector of M_d with eigenvalue λ.

Next we show that the eigenvalues of M_d^* are $\{-d, -d+2, \ldots, d\}$ by induction on d. We use the following identities

$$M_d^* = \begin{bmatrix} & & & & & 0 \\ & & & & & 0 \\ & & \mathbf{M^*_{d-1}} & & & \vdots \\ & & & & & 0 \\ & & & & & n \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ & 1 & 0 & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{bmatrix}$$

$$M_d^* = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ n & & & & & \\ 0 & & & & & \\ 0 & & & \mathbf{M^*_{d-1}} & & \\ \vdots & & & & & \\ 0 & & & & & \end{bmatrix} + \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix}.$$

First note $M_1^* = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ has eigenvalues -1 and 1, establishing the base case. Next we show that if v is an eigenvector of M_{d-1}^* with eigenvalue λ, then $\begin{pmatrix} 0 \\ v \end{pmatrix} - \begin{pmatrix} v \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ v \end{pmatrix} + \begin{pmatrix} v \\ 0 \end{pmatrix}$ are eigenvectors of M_d^* with eigenvalues $\lambda - 1$ and $\lambda + 1$ respectively. Apply the above identities we obtain

$$M_d^* \left(\begin{pmatrix} 0 \\ v \end{pmatrix} - \begin{pmatrix} v \\ 0 \end{pmatrix} \right) = (\lambda - 1) \begin{pmatrix} 0 \\ v \end{pmatrix} + (-\lambda + 1) \begin{pmatrix} v \\ 0 \end{pmatrix} = (\lambda - 1) \left(\begin{pmatrix} 0 \\ v \end{pmatrix} - \begin{pmatrix} v \\ 0 \end{pmatrix} \right)$$

$$M_d^* \left(\begin{pmatrix} 0 \\ v \end{pmatrix} + \begin{pmatrix} v \\ 0 \end{pmatrix} \right) = (\lambda + 1) \begin{pmatrix} 0 \\ v \end{pmatrix} + (\lambda + 1) \begin{pmatrix} v \\ 0 \end{pmatrix} = (\lambda + 1) \left(\begin{pmatrix} 0 \\ v \end{pmatrix} + \begin{pmatrix} v \\ 0 \end{pmatrix} \right).$$

We have shown that $\{-d, -d+2, \dots d\}$ are eigenvalues of M_d. Since M_d has rank $d+1$, this is precisely the set of eigenvalues of M_d. □

Let S be a graph on $n = 2^d$ vertices sampled from M_d. Note that the expected average degree of S is d, and so there approximately $2nd^2$ closed four walks in S that trace trees. In the hypercube, there are $2nd^2$ such walks and an additional nd^2 closed four walks that trace simple cycles. However, the expected number of simple four cycles in S is

$$\mathsf{E}[C_4(S)] = 8 \sum_{i=0}^{n-2} \ell_i \binom{\ell_{i+1}}{2} \ell_{i+2} p_i^2 p_{i+1}^2 = \Theta\left(d^5\right) = o\left(nd^2\right).$$

Therefore the expected number of four walks in S is $(2 + o(1))nd^2$, whereas it is $(3 + o(1))nd^2$ for the hypercubes.