CONNECTIVITY AND GIANT COMPONENT OF STOCHASTIC KRONECKER GRAPHS

MARY RADCLIFFE AND STEPHEN J. YOUNG

Abstract. Stochastic Kronecker graphs are a model for complex networks where each edge is present independently according the Kronecker (tensor) product of a fixed matrix $P \in [0,1]^{k\times k}$. We develop a novel correspondence between the adjacencies in a general stochastic Kronecker graph and the action of a fixed Markov chain. Using this correspondence we are able to generalize the arguments of Horn and Radcliffe on the emergence of the giant component from the case where $k = 2$ to arbitrary k. We are also able to use this correspondence to completely analyze the connectivity of a general stochastic Kronecker graph.

1. Introduction

In many ways the study of random graphs traces its history back to the seminal work of Erdős and Rényi showing that there exists a rapid transition between the regimes of a random graph consisting of many small components, a random graph having one "giant" component, and a random graph being connected [10]. Because of their central role in the history of random graphs these phase transitions have been extensively studied, see for instance [1, 2, 3, 4, 9, 12, 16], among numerous others. We contribute to this ongoing discussion by providing a sharp transition for the emergence of both the giant component and connectivity for the stochastic Kronecker graph, a generalization of the standard Erdős-Rényi binomial random graph model, $\mathcal{G}(n,p)$.

More formally, recall that the Kronecker or tensor product of two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$ is a matrix $A \otimes B = C \in \mathbb{R}^{mp\times nq}$. For $i \in [m], j \in [n], s \in [p]$, and $t \in [q]$ the entry $C_{(i-1)m+s,(j-1)n+t}$ is $A_{ij}B_{st}$, that is

$$A \otimes B = C = \begin{bmatrix} A_{1,1}B & A_{1,2}B & \cdots & A_{1,n}B \\ A_{2,1}B & A_{2,2}B & \cdots & A_{2,n}B \\ \cdots & \cdots & \ddots & \cdots \\ A_{m,1}B & A_{m,2}B & \cdots & A_{m,n}B \end{bmatrix}.$$

Letting $P \in [0,1]^{k\times k}$ be a symmetric matrix, the t^{th}-order stochastic Kronecker graph generated by P is formed by taking t-fold Kronecker product of P, denoted $P^{\otimes t}$, and using this as the probability matrix for a graph with independent edges. That is, each edge $\{i,j\}$ is present independently with probability $P^{\otimes t}_{ij} = P^{\otimes t}_{ji}$.

The stochastic Kronecker graph was originally proposed as model for the network structure of the internet with the property that it could be easily fit to real world data, especially in the case where the generating matrix was $\begin{bmatrix}\alpha & \beta \\ \beta & \gamma\end{bmatrix}$ where $0 < \gamma \leq \beta \leq \alpha < 1$ [14]. As such, there have been several papers analyzing structural properties of the stochastic Kronecker graph when the generating matrix is a 2×2 matrix [14, 15, 17, 19]. Most relevant to this current work are the results of Mahdian and Xu [17] who anaylzed the connectivity, diameter, and the emergence of the giant component with $0 < \gamma \leq \beta \leq \alpha < 1$, and the work of the first author and Horn who analyzed the emergence and size of the giant component for arbitrary $\alpha, \beta, \gamma \in (0,1)$ [19]. In this work we consider the case of an arbitrarily sized generating matrix, and develop necessary and sufficient conditions for the emergence of the giant component and connectivity. The key tool to analyzing these graphs is to tie the structure of the graph to a fixed Markov chain on the underlying generating matrix. Using this underlying structure, one can analyze the graph structure more completely than with traditional tools.

Given a t^{th}-order stochastic Kronecker graph with generating matrix P, we define $W(P)$ to be the weighted graph on $[k]$, where weights are as given in P. We will occasionally refer to W as the underlying graph of

G. We also define the backbone graph of the matrix P, $B(P)$, as the subgraph of $W(P)$ consisting of the edges assigned weight 1. That is, $B(P)$ is a graph on the vertices $[k]$ where $\{i, j\}$ is an edge if and only if $P_{ij} = P_{ji} = 1$. When the matrix P is clear, we will neglect the dependence on P and write simply W and B.

Our primary results can be summarized as follows.

Theorem 1. *Let G be t^{th}-order stochastic Kronecker graph generated by a symmetric matrix $P \in [0,1]^{k\times k}$ which has column sums $c_1 \leq c_2 \leq \cdots \leq c_k$. Let $n = k^t$ be the number of vertices of G.*

(1) *If W is disconnected or bipartite, then the largest component of G has size $\mathcal{O}((k-1)^t) \in o(n)$.*
(2) *If W is connected and non-bipartite and $\prod_i c_i < 1$, then there is some $0 < \alpha < 1$ such that with probability at least $1 - e^{-\Theta(n^\alpha)}$ there are at least $n - \mathcal{O}(n^\alpha)$ isolated vertices in G.*
(3) *If W is connected, non-bipartite, $\prod_i c_i = 1$, and the c_i's are not identically one, then there is a positive constant α such that with probability at least $1 - e^{-\Theta(n^\alpha)}$, the largest component of G has size $\Theta(n)$, that is, G has a giant component.*
(4) *If W is connected, non-bipartite, and $\prod_i c_i > 1$, then there is a positive constant α such that with probability at least $1 - e^{-\Theta(n^\alpha)}$ the largest component of G has size $\Theta(n)$.*
(5) *If W is connected, non-bipartite, and $c_1 < 1$, then there is a positive constant α such that G has at least $\ln(n)^{(1-o(1))\ln\ln(n)}$ isolated vertices with probability at least $1 - \mathcal{O}(n^{-\alpha})$.*
(6) *If W is connected, non-bipartite, $c_1 = 1$, and B has a vertex of degree zero, then there is some positive constant α such that G has at least $\ln(n)^{(1-o(1))\ln\ln\ln(n)}$ isolated vertices with probability at least $1 - \mathcal{O}(n^{-\alpha})$.*
(7) *If W is connected and non-bipartite, $c_1 = 1$, and B has no vertices of degree zero, then there is a constant $\alpha > 0$ such that G is connected with probability at least $1 - e^{-(1-o(1))n^\alpha}$.*
(8) *If W is connected and non-bipartite and $c_1 > 1$, then there is a constant $\alpha > 0$ such that G is connected with probability at least $1 - e^{-(1-o(1))n^\alpha}$.*

We note that item (8) above is typical for the emergence of connectivity; that is, the graph is connected asymptotically almost surely precisely when asymptotically almost surely the minimum degree is at least 1. In fact, taking (5), (6), (7), and (8) together we can see that a stochastic Kronecker graph is connected precisely when the minimum degree is at least 1 with high probability. From this viewpoint, the slightly unnatural seeming condition on the backbone graph B is simply the condition needed to assure that G has no isolated vertices.

The folklore in the study of random graphs asserts that, in general, the giant component should emerge when the average expected degree is 1, see for instance [2, 7, 10, 11]. As the average expected degree in a t^{th}-order stochastic Kronecker graph is $k^{-t}(c_1 + \cdots + c_k)^t$, this suggests that the transition occurs when $\frac{1}{k}(c_1 + \cdots + c_k) > 1$. However, as parts (2) and (4) of Theorem 1 show, the transition actually occurs when $(\prod_i c_i)^{\frac{1}{k}} > 1$. Noting that the expected degrees in stochastic Kronecker graphs follow a multinomial distribution (see Section 2), this condition can be seen as equivalent (asymptotically) to the condition that *median* expected degree is at least one. Thus our results may suggest that the average expected degree is not as deeply connected to giant component as previously thought, because in many of the standard random graph models, such as the Erdős-Rényi random graph, the average and the median expected degree agree. That is, it may be that the median is truly the determining factor for such structures. It is also worth noting that Spencer has conjectured based in part on [5, 6], that the correct intuition is that the emergence of the giant component is tied to the second order average degree [23].

To prove Theorem 1, we will develop several general results on G, and then apply these results to the specific situations above. In particular, we are able to tie the adjacency structure of G to a finite state Markov chain on W. Using this association, we can take advantage of the finite structure of W to build theory a regarding the asymptotically growing structure G.

2. Definitions and Tools

Let G be a stochastic Kronecker graph with generating matrix P as described above. We note that there are multiple means of describing the entries the probability matrix $P^{\otimes t}$ to take advantage of the Kronecker product structure. One point of view that is particularly helpful is to define a bijection $w\colon V(G) \to [k]^t$,

so that each vertex of G is represented by a word of length t in $[k]$. We will often identify the vertex to its corresponding word, and write $v = (v_1, v_2, \ldots, v_t)$. Given an appropriate choice of bijection, for any two vertices u and v, the probability that u and v are adjacent is

$$p_{uv} = \prod_{i=1}^{t} P_{u_i v_i}.$$

That is to say, we take the product of entries of the generating matrix P, where entries correspond to the pairs of components in the words representing u and v.

Let $c_1 \leq c_2 \leq \cdots \leq c_k$ be the column sums of P (note that we can assume these are nondecreasing without loss of generality), and let C be the diagonal matrix of column sums in P. Suppose $w(v)$ has a_1 coordinates equal to 1, a_2 coordinates equal to 2, and so on. It is straightforward to calculate that

$$\mathbb{E}[\deg(v)] = c_1^{a_1} c_2^{a_2} \ldots c_k^{a_k}.$$

Thus it will frequently be of interest to know the number of coordinates in $w(v)$ equal to each symbol in $[k]$. To that end, we define the *signature* of v to be $\sigma(v) = (\sigma_1, \sigma_2, \cdots, \sigma_k)$, where σ_i is the proportion of symbols in $w(v)$ equal to i. For example, if $k = 5$ and $w(v) = 121251$, we would have $\sigma(v) = (\frac{1}{2}, \frac{1}{3}, 0, 0, \frac{1}{6})$. We will denote by $\mathcal{S} = \{(\sigma_1, \ldots, \sigma_k) \mid \sigma_i \geq 0, \sum_i \sigma_i = 1\}$ the space of possible signatures. Often we will establish an underlying signature for a vertex and then take t to infinity; this will generally result in noninteger values for the number of letters of a particular value in $w(v)$. This can be overlooked, however, as rounding to the next integer appropriately will not change the asymptotic features of the vertices, and so we will often assume that a vertex can take any signature.

Let $L = (\ln(c_1), \ln(c_2), \cdots, \ln(c_k))$. We will make frequent use of the simple observation that

$$\ln\left(\mathbb{E}[\deg(v)]\right) = t \left\langle \sigma(v), L \right\rangle,$$

where $\langle \cdot, \cdot \rangle$ represents the standard dot product.

2.1. **Markov chains in G and W.** Let $W^{\otimes t}$ be the weighted complete graph on $V(G)$, with the weight of edge uv equal to $P_{u,v}^{\otimes t}$. Let v be a vertex in $W^{\otimes t}$ with signature $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_k)$. Define $Z^{(v)}$ to be a random variable that takes values in $\mathcal{S}$, where $Z^{(v)}$ is the signature of a randomly chosen neighbor of v according to the probability distribution defined by the weights of the edges. That is,

$$\mathbb{P}(Z^{(v)} = \tau) = \sum_{\sigma(u)=\tau} \frac{P_{u,v}^{\otimes t}}{\deg_{W^{\otimes t}}(v)}.$$

That is to say, $Z^{(v)}$ is the signature of the vertex obtained after taking one step in the uniform random walk on $W^{\otimes t}$.

For each $i \in [k]$, let $X^{(i)}$ be the random variable that takes values in $[k]$, with $\mathbb{P}(X^{(i)} = j) = \frac{P_{ij}}{c_i}$. Note that for $v = (v_1, v_2, \ldots, v_t)$ fixed, we have

$$\mathbb{P}(X^{(v_1)} \times X^{(v_2)} \times \cdots \times X^{(v_t)} = (u_1, u_2, \ldots, u_t)) = \prod_{i=1}^{t} \frac{P_{v_i u_i}}{c_i} = \frac{P_{u,v}^{\otimes t}}{\deg_{W^{\otimes t}}(v)}.$$

Thus we can consider $Z^{(v)}$ as giving the signature of a randomly chosen neighbor of v, chosen according to the product distribution $X^{(v_1)} \times X^{(v_2)} \times \cdots \times X^{(v_t)}$. As the signature is independent of order, for the purposes of analyzing $Z^{(v)}$, we may write this distribution as $(X^{(1)})^{\sigma_1 t} \times (X^{(2)})^{\sigma_2 t} \times \cdots \times (X^{(k)})^{\sigma_k t}$. Therefore, for all $i \in [k]$, letting $Z_i^{(v)}$ be the i^{th} component of the signature $Z^{(v)}$, we have

$$\mathbb{E}\left[Z_i^{(v)}\right] = \frac{1}{t} \sum_j (\sigma_j t) \mathbb{P}(X^{(j)} = i) = \sum_j \sigma_j \frac{P_{ij}}{c_j}.$$

On the other hand, let $M = C^{-1}P$, the transition probability matrix for the uniform random walk on W and notice that the matrix product σM has i^{th} coordinate

$$\left((\sigma_1, \sigma_2, \ldots, \sigma_k)M\right)_i = \sum_{j=1}^{k} \sigma_j M_{ij} = \sum_{j=1}^{k} \sigma_j \frac{P_{ij}}{c_i} = \mathbb{E}\left[Z_i^{(v)}\right]$$

Thus, $\sigma M = \mathbb{E}\left[Z^{(v)}\right]$.

Therefore, we can think of the distribution of a random walk on W as the expected signature of a vertex in a random walk on $W^{\otimes t}$. Let $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ be the stationary distribution of the random walk on W, so $\pi M = \pi$. It is a simple exercise to verify that $\pi_i = \frac{c_i}{\sum_j c_j}$. We will show in Section 3 that the collection of signatures close to π will in fact, asymptotically almost surely, form a connected subgraph in G, and further, by leveraging theconvergence of the Markov chain on W, we can assure a giant component.

2.2. **Tools and Notation.** Given a stochastic Kronecker graph G generated by P, let A be the adjacency matrix of G and D the diagonal matrix of degrees in G. The stochastic Kronecker graph is defined precisely so that the expected adjacency matrix $\bar{A} = P^{\otimes t}$, and the expected degree matrix $\bar{D} = C^{\otimes t}$. We will sometimes use the notation $P^{\otimes t}_{u,v}$ to refer to the $w(u), w(v)$ position in $P^{\otimes t}$, where we index the matrix by the ordered words obtained via the Kronecker product. At times we will wish to emphasize the graph structure of $P^{\otimes t}$ and thus will refer to it as $W^{\otimes t}$.

Among our key tools will the the following theorem from Chung and the first author [18] that gives spectral concentration in the normalized Laplacian of a general random graph.

Theorem 2 ([18])**.** *Let G be a random graph with independent edges generated according to the matrix $\mathcal{P}$ and let A be the associated adjacency matrix. Let D be the diagonal matrix of expected degrees and let δ denote the minimum expected degree.If $\delta \geq 3\ln\left(\frac{4n}{\epsilon}\right)$, then with probability at least $1-\epsilon$ we have that, for all i*

$$\left|\lambda_i\left(\mathcal{L}\left(G\right)\right) - \lambda_i\left(I - D^{-1/2}\mathcal{P}D^{-1/2}\right)\right| \leq 3\sqrt{\frac{3\ln\left(\frac{4n}{\epsilon}\right)}{\delta}}.$$

We also make use of standard tools in spectral graph theory, chief among them the Cheeger inequality. For two sets S, T of vertices in a graph G, define $e_G(S,T)$ to be the number of edges (or, in a weighted graph, the total weight of edges) for which one endpoint is in S and the other in T. Define $\mathrm{Vol}_G(S) = \sum_{v\in S}\deg(v)$. When the underlying graph is clear, we drop the subscript G in the notation.

The Cheeger constant of a set S with $\mathrm{Vol}(S) \leq \frac{1}{2}\mathrm{Vol}(G)$ is defined to be $h(S) = {}^{e(S,V\setminus S)}/_{\mathrm{Vol}(S)}$ and Cheeger constant of G is

$$h_G = \min_{\substack{S\subset V \\ \mathrm{Vol}(S)\leq\frac{1}{2}\mathrm{Vol}(G)}} h(S).$$

The spectrum of a graph is related to the Cheeger constant via the Cheeger Ineqaulity [21, 22].

Cheeger Inequality. *For G any graph, let λ_1 be the smallest nonzero eigenvalue of $\mathcal{L}(G)$. Then*

$$\frac{1}{2}h_G^2 \leq \lambda_1 \leq 2h_G.$$

As we will frequently be discussing Markov chains, we will pass regularly between considering row vectors and column vectors. We will always treat the signature of a vertex v as a row vector, as well as the vector L. The all-ones vector, $\mathbb{1}$, will be considered a row vector as well. However, eigenvectors of a matrix are typically assumed to be right eigenvectors, and are thus column vectors. Any other usages should be made clear by context.

In order to understand the rate of convergence of a Markov chain we will use the relative pointwise distance. If π is the limiting distribution of the Markov chain, the relative pointwise distance of a distribution σ from π is

$$\Delta_{RP}(\sigma) = \max_i \frac{|\sigma_i - \pi_i|}{\pi_i}.$$

As we are interested in an overall rate of convergence we define

$$\Delta(s) = \sup_{\sigma\in\mathcal{S}} \Delta_{RP}\left(\sigma M^s\right).$$

It is well know that the rate of decay of the relative pointwise distance can be controlled by the spectral information of the Markov chain as given in the following theorem, see for instance [8].

Theorem 3. *Let $1 = \lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{n-1}$ by the eigenvalues of the transition probability matrix of a uniform random walk on a connected, non-bipartite (weighted) graph G. Set $\lambda = \max\{|1-\lambda_1|, |\lambda_{n-1}-1|\}$. For any*

$$s > \frac{1}{\lambda}\ln\left(\frac{\mathrm{Vol}(G)}{\epsilon\delta_G}\right),$$

we have $\Delta(s) < \epsilon$, where δ_G denotes the minimum degree in G.

The phrase *asymptotically almost surely* in this paper will always refer to asymptotics with respect to t, unless otherwise noted. The norm $\|v\|$ will refer to the ℓ_∞-norm unless otherwise noted.

3. Key Results

To prove the thresholds for connectivity and emergence of the giant component in a stochastic Kronecker graph G (Theorem 1, items (4) and (8)), we will use the following structure. First, we show that G contains a small set of vertices that is connected asymptotically almost surely, in particular, those vertices that are close to stationarity under the Markov chain described in Section 2.1. We shall refer to this set as the "connected core" of the graph. Although this will not be enough vertices to form a giant component, we can then show that under certain conditions, a positive fraction of the vertices in G can be connected by a path to the connected core. The thresholds given are precisely those conditions needed to ensure that a positive fraction of the vertices exhibit this behavior. In retrospect, the arguments used by Horn and the first author in [19] to show the emergence of the giant component in the case where the generating matrix is 2×2 can be viewed as a special case of our technique. Specifically, as the underlying Markov chain has only two states, the degree of each vertex is controlled by a single parameter, which significantly simplifies the argument. As a consequence, the authors in [19] were able to analyze the giant component directly via counting techniques, without appealing to the underlying Markov chain.

In this section, we develop much of the underlying structure in G via the random walk on W. We begin with some elementary observations on the vertex degrees in G and $W^{\otimes t}$.

Lemma 4. *Let v be a vertex with signature σ in a t^{th}-order stochastic Kronecker graph G, such that $\langle\sigma, L\rangle > 0$. Let $d = e^{\langle\sigma,L\rangle}$. For any $\delta > 0$, we have*

1. *v has at least $d^t(1-2ke^{-2\delta^2 t})$ neighbors in $W^{\otimes t}$ with signature τ such that $\left\|\tau - \mathbb{E}\left[Z^{(v)}\right]\right\| \leq \delta$.*
2. *with probability at least $1-\exp(-\frac{d^t}{8}(1-2ke^{-2\delta^2 t}))$, v has at least $\frac{1}{2}d^t(1-2ke^{-2\delta^2 t})$ neighbors in G with signature τ such that $\left\|\tau - \mathbb{E}\left[Z^{(v)}\right]\right\| \leq \delta$.*

Proof. By the Hoeffding inequality, we have that for any i,

$$\mathbb{P}\left(t\left|Z_i^v - \mathbb{E}\left[Z_i^{(v)}\right]\right| > \delta t\right) \leq 2e^{-2\delta^2 t}$$

for any $\delta > 0$. Therefore, by the union bound, we have

$$\mathbb{P}\left(\exists i \in [k] \text{ such that } t\left|Z_i^{(v)} - \mathbb{E}\left[Z_i^{(v)}\right]\right| > \delta t\right) \leq 2ke^{-2\delta^2 t}.$$

This verifies item (1).

For item (2), note that by (1), we have that expected number of neighbors of v with signature τ in the desired range desired is at least $d^t(1-2ke^{-2\delta^2 t})$. By Chernoff bounds, then, with probability at least $1-\exp(-\frac{d^t}{8}(1-2ke^{-2\delta^2 t}))$, we have at least $\frac{1}{2}d^t(1-2ke^{-2\delta^2 t})$ neighbors with such a signature τ. □

As an immediate corollary of this result we have the following.

Corollary 5. *Let v be a vertex with σ in a t^{th}-order stochastic Kronecker graph G, such that $\langle\sigma, L\rangle > 0$. Let $d = e^{\langle\sigma,L\rangle} > 1$. With probability at least $1-e^{-\frac{d^t}{12}}$, v has at least $\frac{d^t}{3}$ neighbors u with $\|\sigma(u) - \sigma M\| \leq \sqrt{\frac{\ln(6k)}{2t}}$.*

Recall from Section 2.1 that $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ is the stationary distribution of the random walk on W, with $\pi_i = \frac{c_i}{\mathrm{Vol}(W)}$ for all i. Given $\epsilon > 0$, define $\mathcal{S}_\epsilon = \{v \in G \mid \forall i \in [k], \sigma_i(v) > (1-\epsilon)\pi_i\}$. We will show that under appropriate conditions, this set of vertices $\mathcal{S}_\epsilon$ is connected asymptotically almost surely, forming the small connected core described above. To do this, we will show that vertices in $\mathcal{S}_\epsilon$ have exponentially large degree in t, and then use Theorem 2 to show the first eigenvalue in $\mathcal{S}_\epsilon$ is bounded away from zero. We first must address the degree of vertices in $\mathcal{S}_\epsilon$. To that end, we have the following Lemma:

Lemma 6. *Let G be a t^{th}-order stochastic Kronecker graph generated by P and let $\epsilon > 0$ be fixed. For sufficiently large t there is a constant c, depending only on P and ϵ, such that for all $v \in \mathcal{S}_\epsilon$, $\mathbb{P}\big(Z^{(v)} \in \mathcal{S}_\epsilon\big) \geq c$.*

To prove this Lemma, we make use of the following standard observation about binomial random variables.

Observation 7. *Let $\alpha_1 > \alpha_2$ be fixed constants and let $p \in (0,1)$. There exists constants c and n_0, depending on α_1, α_2, and p such that if $n > n_0$, then*

$$\mathbb{P}(\text{Bin}\,(n,p) \in [np - \alpha_1\sqrt{np}, n - \alpha_2\sqrt{np}]) > c.$$

Proof of Lemma 6. Let v be an arbitrary vertex in $\mathcal{S}_\epsilon$. Consider a collection of independent, identically distributed random variables, $X_1, \ldots, X_t$, taking on values in $\{1, \ldots, k\}$ each with probability p_i, where $p_i \geq p > 0$ for all i. Let Z_i be the count of the number of i's in these variables, that is, $Z_i = \sum_j \mathbb{1}_{X_j = i}$. Let $\mathcal{E}_i$ be the event that $p_i t - 2c\sqrt{t} \leq Z_i \leq p_i t - c\sqrt{t}$. We then have that, for all $j \neq i$,

$$\begin{aligned}
\mathbb{E}[Z_j \mid \mathcal{E}_i] &\geq \left(t - \left(p_i t - c\sqrt{t}\right)\right) \frac{p_j}{1 - p_i} \\
&= \left((1-p_i)t + c\sqrt{t}\right) \frac{p_j}{1-p_i} \\
&= p_j t + \frac{c p_j}{1 - p_i}\sqrt{t} \\
&\geq p_j t + cp\sqrt{t}.
\end{aligned}$$

To apply this observation to the context of $Z^{(v)}$ we first consider the unweighted graph W' on $[k]$ where $i \sim j$ if and only if there is an unweighted walk of length 2 between i and j in W. Since W is non-bipartite, W' is connected and thus there exists a breadth-first traversal of W'. We note that by the definition of $\mathcal{S}_\epsilon$, for every i we have $(\sigma M)_i \geq (1-\epsilon)\pi_i$. Further, by the pigeonhole principle, there is some index i such that $(\sigma M)_i \geq \pi_i(1-\epsilon) + \frac{\epsilon}{k}$. Let s_1 be one such index and let $s_1, \ldots, s_k$ be a breadth-first traversal of W' starting at s_1.

Recall that we may analyze $Z^{(v)}$ from the point of view of the product distribution $\left(X^{(1)}\right)^{\sigma_1 t} \times \cdots \times \left(X^{(k)}\right)^{\sigma_k t}$ where each $X^{(i)}$ is an independent random variable that takes values in the set of neighbors of i in W. Let the random variables Z_{ij} be the number of times that $X^{(i)}$ takes on the value j. We note that we can ignore the indices that $X^{(i)}$ can not take on, and so define $p_i = \min_{j, p_{ij} \neq 0} \frac{p_{ij}}{c_i}$. We recursively define the events $\mathcal{E}_1, \ldots, \mathcal{E}_k$ as follows. The event $\mathcal{E}_1$ is the event that for all $u \sim_W s_1$, $\mathbb{E}[Z_{us_1}] - 2\alpha_1\sqrt{t} \leq Z_{us_1} \leq \mathbb{E}[Z_{us_1}] - \alpha_1\sqrt{t}$. For all $1 < i \leq k$ the event $\mathcal{E}_i$ is the event that for all $u \sim_W s_i$,

$$\mathbb{E}\big[Z_{us_i} \mid \cap_{j=1}^{i-1}\mathcal{E}_j\big] - 2\alpha_i\sqrt{t} \leq Z_{us_i} \leq \mathbb{E}\big[Z_{us_i} \mid \cap_{j=1}^{i-1}\mathcal{E}_j\big] - \alpha_i\sqrt{t},$$

where the α_i's are fixed constant to be chosen later. We note that by Observation 7 that each of these events occurs with positive probability, thus it suffices to show that $\cap_{i=1}^k \mathcal{E}_i$ is contained in the event $Z^{(v)} \in \mathcal{S}_\epsilon$.

For sufficiently large t the event $\mathcal{E}_1$ assures that $\left(Z^{(v)}\right)_{s_1} \geq (1-\epsilon)\pi_{s_1}$ by the choice of s_1, specifically that $\mathbb{E}\left[Z^{(v)}_{s_1}\right] \geq \pi_{s_1} + \frac{\epsilon}{k}$.

Since the sequence s_i is a breadth-first search of W', we have that for all $i > 1$, there exists index $j < i$ such that $s_i \sim_{W'} s_j$. Thus there is some vertex u that is a neighbor to both s_i and s_j in W. Now consider the effect of the conditioning on the event $\mathcal{E}_j$ on Z_{us_i}. By the above calculation and the definition of $\mathcal{E}_j$ we have that implies that $\mathbb{E}\big[Z_{us_i} \mid \cap_{j=1}^{i-1}\mathcal{E}_j\big] \geq \mathbb{E}[Z_{us_i}] + \alpha_{i-1}p_u\sqrt{t} \geq \mathbb{E}[Z_{us_i}] + \alpha_{i-1}p_{\min}\sqrt{t}$ where $p_{\min} = \min_{i \in [k]} p_i$. Furthermore, this gives that $t\mathbb{E}\left[Z^{(v)}_{s_i} \mid \cap_{j=1}^{i-1}\mathcal{E}_j\right] \geq (1-\epsilon)\pi_{s_i}t + \alpha_{i-1}p_{\min}\sqrt{t}$. Thus choosing $\alpha_i = \left(\frac{2k}{p_{\min}}\right)^{k-i}$ suffices to assure that the event $\cap_{i=1}^k \mathcal{E}_i$ is contained in $\mathcal{S}_\epsilon$, as desired. □

Theorem 8. *Let G be a t^{th}-order stochastic Kronecker graph generated by a matrix $P \in [0,1]^{k \times k}$ such that W is connected and non-bipartite. Further supposed that $\sum_i c_i \ln(c_i) > 0$ and fix*

$$0 < \epsilon < \frac{\sum c_i \ln(c_i)}{\sum c_i \ln(c_i) - \text{Vol}(W)\ln(c_1)}.$$

Let H be the subgraph of G induced by $\mathcal{S}_\epsilon$. For t sufficiently large, there is a constant $d > 1$, depending on P and ϵ, such that H is connected with diameter $\mathcal{O}(\log|\mathcal{S}_\epsilon|)$ with probability at least $1 - e^{\Theta\left(d^t\right)}$.

Notice that the bound on ϵ is always positive (or infinite), since $c_1 \leq c_i$ for all i, so $\mathrm{Vol}(W)\ln c_1 = \sum c_i \ln(c_1) \leq \sum c_i \ln(c_i)$.

Proof. We will proceed by showing that the graph H has an asymptotically constant spectral gap and thus by standard results in spectral graph theory, $\mathcal{S}_\epsilon$ is connected with diameter $\mathcal{O}(\ln(|\mathcal{S}_\epsilon|))$.

Recall that the expected degree of a vertex with signature σ is $(c_1^{\sigma_1}\cdots c_k^{\sigma_k})^t$ and thus any vertex $v \in \mathcal{S}_\epsilon$ has expected degree at least

$$c_1^{\epsilon t}\left(c_1^{\pi_1}\cdots c_k^{\pi_k}\right)^{(1-\epsilon)t} = \left(c_1^{\epsilon}\left(c_1^{c_1}\cdots c_k^{c_k}\right)^{\frac{1-\epsilon}{\mathrm{Vol}(W)}}\right)^t = d^t,$$

where

$$d = c_1^{\epsilon}\left(c_1^{c_1}\cdots c_k^{c_k}\right)^{\frac{1-\epsilon}{\mathrm{Vol}(W)}}.$$

We note that by the restriction on ϵ,

$$\begin{aligned}\ln(d) &= \epsilon\ln(c_1) + \frac{1-\epsilon}{\mathrm{Vol}(W)}\sum_i c_i\ln(c_i)\\ &= \frac{1}{\mathrm{Vol}(W)}\sum_i c_i\ln(c_i) + \epsilon\left(\ln(c_1) - \frac{1}{\mathrm{Vol}(W)}\sum_i c_i\ln(c_i)\right)\\ &> 0,\end{aligned}$$

and thus $d > 1$. This implies that every vertex in $\mathcal{S}_\epsilon$ has expected degree exponentially increasing with t.

Let $\overline{H}$ be the subgraph of $W^{\otimes t}$ induced by $\mathcal{S}_\epsilon$, so the weight of each edge in $\overline{H}$ is the expectation of that edge appearing in H. Now, by Lemma 6, there is some constant c such that for every vertex v in $\overline{H}$ we have $\deg_{\overline{H}}(v) \geq cd^t$. Now for any positive constant δ, there exists some small positive constant c' such that

$$\frac{27\ln\left(\frac{4|\mathcal{S}_\epsilon|}{e^{-c'd^t}}\right)}{cd^t} \leq \frac{27\ln\left(\frac{4k^t}{e^{-c'd^t}}\right)}{cd^t} = \frac{27\left(t\ln(k)+\ln(4)-c'd^t\right)}{cd^t} = o(1) + \frac{27c'}{c} \leq \delta^2,$$

and thus, by Theorem 2, in order to complete the proof it suffices to show that $\overline{H}$ has constant spectral gap.

To determine the spectral gap in $\overline{H}$, we use Cheeger's inequality. Let $X \subset \mathcal{S}_\epsilon$ with $\mathrm{Vol}_{\overline{H}}(X) < \frac{1}{2}\mathrm{Vol}_{\overline{H}}(\mathcal{S}_\epsilon)$. Note that

$$h_{\overline{H}}(X) = \frac{e_{\overline{H}}(X, \mathcal{S}_\epsilon\backslash X)}{\mathrm{Vol}_{\overline{H}}(X)} \geq \frac{e_{W^{\otimes t}}(X, V\backslash X)}{\frac{1}{c}\mathrm{Vol}_{W^{\otimes t}}(X)} = c\, h_{W^{\otimes t}}(X),$$

where the constant c is the constant provided by Lemma 6. Thus, we have

$$\begin{aligned}h_{\overline{H}} &= \min_{\substack{X\subset\mathcal{S}_\epsilon\\ \mathrm{Vol}(X)<\frac{1}{2}\mathrm{Vol}(\mathcal{S}_\epsilon)}} h_{\overline{H}}(X)\\ &\geq c \min_{\substack{X\subset\mathcal{S}_\epsilon\\ \mathrm{Vol}(X)<\frac{1}{2}\mathrm{Vol}(\mathcal{S}_\epsilon)}} h_{W^{\otimes t}}(X)\\ &\geq c\, h_{W^{\otimes t}}.\end{aligned}$$

Now, let $M_1 = C^{-1/2}PC^{-1/2}$ and let $1 = \mu_0 \geq \mu_1 \geq \cdots \geq \mu_{k-1}$ be the eigenvalues of M_1. Note that $I - M_1$ is the Laplacian matrix for W, and as W is connected and non-bipartite, $-1 < \mu_{k-1} \leq \mu_1 < 1$. Now, $\mathcal{L}(W^{\otimes t}) = I - M^{\otimes t}$, and thus has eigenvalues $1 - \mu_{a_1}\mu_{a_2}\cdots\mu_{a_t}$, where $a_1, a_2, \ldots, a_t \in [k-1]\cup\{0\}$. Hence, the smallest nonzero eigenvalue of $\mathcal{L}\left(W^{\otimes t}\right)$ is $1-\mu_1$, which occurs with multiplicity t. Thus by Cheeger's inequality, $h_{W^{\otimes t}} \geq \frac{1-\mu_1}{2}$.

Therefore, combining these results we have

$$\lambda_1\left(\overline{H}\right) \geq \frac{1}{2}h_{\overline{H}}^2 \geq \frac{c}{2}h_{W^{\otimes t}}^2 \geq \frac{c^2}{8}(1-\mu_1)^2.$$

Hence $\lambda_1\left(\overline{H}\right)$ is bounded below by a constant and $\overline{H}$ has constant spectral gap, as desired. □

This establishes that the graph G contains a small connected core asymptotically almost surely provided $\sum c_i \ln c_i > 0$. We now turn our attention to the second half of our fundamental structure. Here we wish to determine which vertices will be connected by a path to the connected core. To that end, define

$\Sigma_\nu = \{v \in V(G) \mid \langle \sigma(v)M^s, L\rangle \geq \nu \text{ for all } s \geq 0\}$. We wish to show that any vertex in Σ_ν may be connected by a path to $\mathcal{S}_\epsilon$ asymptotically almost surely.

Theorem 9. *Let G be a t^{th}-order stochastic Kronecker graph generated by a matrix $P \in [0,1]^{k\times k}$ such that W is connected and non-bipartite. Fix $0 < \epsilon, \nu$. Let λ be the spectral gap of W and let $s = \left\lceil \frac{1}{\lambda} \ln\left(\frac{2\operatorname{Vol}(W)}{c_1\epsilon}\right)\right\rceil$. For t sufficiently large, any vertex $v \in \Sigma_\nu$ is connected to $\mathcal{S}_\epsilon$ by a path of length at most s with probability at least $1 - se^{-e^{\nu t - \Theta(\sqrt{t})}}$.*

Proof. Let $v \in \Sigma_\nu$. Define $v_0 = v$ and for each $1 \leq i \leq s$, let v_i be a neighbor of v_{i-1} such that $\|\sigma(v_i) - \sigma(v_{i-1})M\| \leq \sqrt{\frac{\ln(6k)}{2t}}$ (if such a neighbor exists). For $1 \leq i \leq s$ define $\eta_i = \sigma(v_i) - \sigma(v_{i-1}M)$. Now, we note that if such as sequence exists, then

$$\left\|\sigma(v_j) - \sigma(v)M^j\right\| \leq \left\|\sum_{i=1}^{j} \eta_i M^{j-i}\right\| \leq \sum_{i=1}^{j} \left\|\eta_i M^{j-i}\right\| \leq \sum_{i=1}^{j} \|\eta_i\|_1 \leq \sum_{i=1}^{j} k\sqrt{\frac{\ln(6k)}{2t}} = jk\sqrt{\frac{\ln(6k)}{2t}},$$

and further

$$\langle v_j, L\rangle \geq \langle v_0, L\rangle - jk\sqrt{\frac{\ln(6k)}{2t}}\,\|L\|_1 \geq \nu - jk\sqrt{\frac{\ln(6k)}{2t}}\,\|L\|_1 .$$

Thus, since s is a fixed constant, we have that by Corollary 5 for sufficiently large t such a sequence fails to exist with probability at most

$$se^{-\frac{\left(e^{\nu - sk\sqrt{\frac{\ln(6k)}{2t}}\|L\|_1}\right)^t}{12}} = se^{-\frac{e^{\nu t - \Theta(\sqrt{t})}}{12}} = se^{-e^{\nu t - \Theta(\sqrt{t})}}$$

It now suffices to show that $v_s \in \mathcal{S}_\epsilon$.

By the choice of s and Theorem 3, we know that

$$\left|\frac{(\sigma(v)M^s)_i - \pi_i}{\pi_i}\right| \leq \frac{\epsilon}{2},$$

and thus $(\sigma(v)M^s)_i \geq (1 - \frac{\epsilon}{2})\pi_i$. But then as $|(v_s)_i - (\sigma(v)M^s)_i| \leq sk\sqrt{\frac{\ln(6k)}{2t}}$ we have that for sufficiently large t, $v_s \in \mathcal{S}_\epsilon$. □

4. Small Components

We now turn to the case that the stochastic Kronecker graph has only small components, that is, the largest component is of size at most $o(n) = o(k^t)$. These correspond to items (1) and (2) in Theorem 1. The first of these result follows from standard results on the component sizes of (non-stochastic) Kronecker graphs which we include in the following lemma for completeness.

Lemma 10. *If H is a disconnected or bipartite graph on k vertices, then the largest component of $H^{\otimes t}$ has size $\mathcal{O}((k-1)^t)$.*

Proof. First, suppose H is not connected. Let $v = (v_1, v_2, \ldots, v_t)$ be a vertex in $H^{\otimes t}$. Now for any neighbor $u = (u_1, u_2, \ldots, u_t)$ of v each coordinate u_i must be adjacent to v_i in H and hence in the same component as v_i. Thus, the size of the component containing v is at most the product of the sizes of the components in H of the vertices v_i. Since H is disconnected the largest component in H has size at most $k-1$ and thus the largest component in $H^{\otimes t}$ has size at most $(k-1)^t$.

Now, suppose H is a connected bipartite graph with bipartition (A, B) and again consider a vertex $v = (v_1, v_2, \ldots, v_t)$ and a neighbor u of v, with $u = (u_1, u_2, \ldots, u_n)$. Now since v_i and u_i are adjacent in H, they are on different sides of the bipartition (A, B). Thus the component v and u are in is bipartite with u and v on different sides of the bipartition. Furthermore, the side of the bipartition containing v is $|A|^{\{i:v_i\in A\}}\,|B|^{\{j:v_j\in B\}}$. Thus for all $0 \leq i \leq t$ there are $\binom{t}{i}$ components of $H^{\otimes t}$ of size $|A|^i\,|B|^{t-i} + |A|^{t-i}\,|B|^i$. It is worth noting that this size is symmetric and so that components counted for a given i are also counted for $t-i$. Now maximizing $|A|^i\,|B|^{t-i} + |A|^{t-i}\,|B|^i$ over the choice of i, we have the largest component occurs where either $i = 0$ or $i = t$. As $|B| = k - |A|$, we maximize with respect to $|A|$ to obtain that the largest of component of $H^{\otimes t}$ has size at most $(k-1)^t + 1$ for $k > 1$. □

This lemma resolves item (1) in Theorem 1 as it implies that the underlying graph for $P^{\otimes t}$ is disconnected with small component sizes.

Theorem 11. *Let G be a t^{th}-order stochastic Kronecker graph generated by $P \in [0,1]^{k\times k}$ with column sums $c_1 \leq \cdots \leq c_k$. If W is connected, non-bipartite, and $\prod_i c_i < 1$, then there some $0 < \delta < 1$ such that with probability at least $1 - e^{-\frac{n^\delta}{3}}$ there are at least $n - \mathcal{O}(n^\delta)$ isolated vertices in G.*

Proof. As $\prod_i c_i < 1$ we have that $\sum_i \ln(c_i) = -\epsilon k < 0$. Let α be a solution to

$$\alpha = \frac{2(\epsilon-\alpha)^2}{(\ln(c_k)-\ln(c_1))^2}$$

in the interval $[0,\epsilon]$. Such an α exists as α and $\frac{2(\epsilon-\alpha)^2}{(\ln(c_k)-\ln(c_1))^2}$ are continuous functions, $0 < \frac{2\epsilon^2}{(\ln(c_k)-\ln(c_1))^2}$, and $\epsilon > 0$. Let $\delta = 1 - \frac{\alpha}{\ln(k)}$. Let $X = X_1 + \cdots + X_t$ where each X_i takes values independently uniformly from $\{\ln(c_1), \ldots, \ln(c_k)\}$. Note that X can be thought of as the natural logarithm of the expected degree of a vertex of G chosen uniformly at random. Now by Hoeffding bounds we have that

$$\mathbb{P}(X \geq -\alpha t) = \mathbb{P}(X + \epsilon t \geq (\epsilon - \alpha)t) \leq e^{-\frac{2(\epsilon-\alpha)^2}{(\ln(c_k)-\ln(c_1))^2}t} = e^{-\alpha t}.$$

Thus there are at most $k^t e^{-\alpha t} = n^\delta$ vertices of G with expected degree smaller than $e^{-\alpha t}$. The sum of the expected degrees of vertices with expected degree larger than $e^{-\alpha t}$ is at most $k^t e^{-\alpha t} = n^\delta$. Thus by Chernoff bounds with probability at least $1 - e^{-\frac{n^\delta}{3}}$ there are at most $2n^\delta$ edges incident to vertices with expected degree at most $e^{-\alpha t}$. Combining this with the vertices with expected degree at least $e^{-\alpha t}$ we have that there are at most $3n^\delta$ non-isolated vertices in G. □

The preceding theorem resolves item (2) in Theorem 1.

5. Giant Components

We now turn our attention to proving item (4) in Theorem 1. To prove this result, we will use the structure outlined in Section 3, and in particular, Theorems 8 and 9 regarding the existence of a connected core of vertices and the vertices that can be connected by a path to S_ϵ. In order to apply these theorems, however, we must verify that the conditions are met. We thus begin with several additional lemmas addressing the case that $\prod_i c_i > 1$.

Lemma 12. *Let $0 < c_1 \leq \cdots \leq c_k$ be such that $\prod_i c_i \geq 1$. Then $\sum_i c_i \ln(c_i) \geq 0$ with equality if and only if the c_i's are identically 1.*

Proof. Define $\delta_j = c_j - c_{j-1} \geq 0$, where c_0 is defined to be 0 and define $s_j = \sum_{i=j}^k \ln(c_i)$. As $\sum_i c_i \ln(c_i) = \sum_i \delta_i s_i$, and all the $\delta_i \geq 0$, it suffices to show that $s_i \geq 0$. We note that since the c_i's are increasing and $\ln(\cdot)$ is a monotonically increasing function $0 \leq \sum_i \ln(c_i) \leq \frac{j-1}{k-j+1} s_j + s_j$, and thus $s_j \geq 0$ for all j.

We note that if $\prod_i c_i > 1$, then the previous argument implies that $\sum_i c_i \ln(c_i) > 0$. Thus suppose that $\prod_i c_i = 1$ and yet the c_i's are not identically 1. As this implies that $c_k > 1$ and $c_1 < 1$, there is some minimal j such that $c_j > 1$. But then as $c_{j-1} \leq 1$, $\delta_j > 0$ and $s_j = \sum_{i=j}^k \ln(c_i) \geq (k-j+1)\ln(c_j) > 0$, we have that $\sum_i c_i \ln(c_i) > 0$, as desired. □

Lemma 13. *Let P be a symmetric matrix in $[0,1]^{k\times k}$ with non-identical column sums $0 < c_1 \leq \cdots \leq c_k$. Further suppose that the associated weighted graph W is connected and non-bipartite. Let f be a strictly monotonically increasing function on $\mathbb{R}^+$ and let L be the vector $(f(c_1), \ldots, f(c_k))$. If M is the transition matrix for the uniform random walk on W, then $\langle \mathbb{1}M^s, L\rangle > \langle \mathbb{1}, L\rangle$ for all $s \geq 1$.*

Proof. We first note that $M = C^{-1}P$ and consider

$$\begin{aligned}
\langle \mathbb{1}M, L\rangle - \langle \mathbb{1}, L\rangle &= \left\langle \mathbb{1}C^{-1}P, L\right\rangle - \langle \mathbb{1}, L\rangle \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{P_{ij}}{c_i} L_j - \sum_{j=1}^{k} L_j \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{P_{ij}}{c_i} L_j - \sum_{j=1}^{k}\sum_{i=1}^{k} \frac{P_{ij}}{c_j} L_j \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} \left(\frac{P_{ij}}{c_i} - \frac{P_{ij}}{c_j}\right) L_j \\
&= \sum_{c_i > c_j} P_{ij}\left(\frac{1}{c_j} - \frac{1}{c_i}\right)(L_i - L_j)
\end{aligned}$$

Note that as f is monotonically increasing, $L_j - L_i > 0$ and $\frac{1}{c_j} - \frac{1}{c_i} > 0$ for $c_i > c_j$. Further, as W is connected, $P_{ij} > 0$ for some i and j with $c_i \neq c_j$, giving that $\langle \mathbb{1}M, L\rangle - \langle \mathbb{1}, L\rangle > 0$.

To complete the proof it would suffice to show that M^s is the transition probability matrix for the uniform random walk on some connected, non-bipartite graph with the same degree sequence as W. To that end, fix some $s \geq 2$ and note that $M^s = C^{-1}\left(PC^{-1}\right)^{s-1}P$, and so let $P' = \left(PC^{-1}\right)^{s-1}P$. It is clear that P' is symmetric and has the desired column sums, thus it suffices to show that the associated graph W' is connected and non-bipartite. We note that $P'_{ij} > 0$ if and only if there is a length s walk between i and j in W. We note that if s is odd, then the edges present in W' are a superset of the edges in W, and thus W' is connected and non-bipartite.

Thus suppose s is even and let C be an odd length cycle in W. Consider the walk in W' formed by starting at vertex v and traversing the cycle C in steps of length s. As s is even and the length of the cycle is odd, it will take an odd number of steps in W' to return to the vertex v. Thus, there is a closed walk in W' of odd length and hence W' is non-bipartite. We note that as s is even W' contains self-loops at all vertices and edges between pairs of vertices that are connected by a walk of length 2. Thus in order to show that W' is connected it suffices to show that there is an even length walk between any two vertices in W. For any two distinct vertices u and v in W such a walk can be constructed by taking a walk from each vertex to the odd cycle C and then traversing C in both directions. As C is an odd cycle, these two traversals will have opposite parity, and thus one of those walks will have even length. □

These two Lemmas immediately give part (4) of our main theorem, as follows.

Theorem 14. *Let G be a t^{th}-order stochastic Kronecker graph generated by a matrix $P \in [0,1]^{k\times k}$ such that W is connected and non-bipartite. If $\prod_i c_i > 1$, then there are constants $s, d > 1$, depending only on P, such that for sufficiently large t, G has a giant component with probability at least $1 - sk^t e^{-\Theta\left(d^t\right)}$.*

Proof. By Lemma 12, we have that $\sum_i c_i \ln(c_i) > 0$. Now fix

$$0 < \epsilon = \frac{\sum_i c_i \ln(c_i)}{2\sum_i c_i \ln(c_i) - 2\ln(c_1)\operatorname{Vol}(W)} < \frac{\sum_i c_i \ln(c_i)}{\sum_i c_i \ln(c_i) - \ln(c_1)\operatorname{Vol}(W)}.$$

By Theorem 8, there is some constant $d_1 > 1$ which depends only on P such that $\mathcal{S}_\epsilon$ is connected with probability at least $1 - e^{-\Theta\left(d_1^t\right)}$.

Fix some positive constant c. Let v be an arbitrary vertex such that $\left\|\sigma(v) - \frac{1}{k}\mathbb{1}\right\| \leq \frac{c}{\sqrt{t}}$ and let $\eta_v = \sigma(v) - \frac{1}{k}\mathbb{1}$. Noting that $\langle \mathbb{1}, L\rangle = \ln\left(\prod_i c_i\right) > \ln(1) = 0$, we have that for sufficiently large t and all $s \geq 0$,

$$\begin{aligned}\langle \sigma(v)M^s, L\rangle &= \left\langle \left(\frac{1}{k}\mathbb{1} + \eta_v\right) M^s, L \right\rangle \\ &= \frac{1}{k}\langle \mathbb{1}M^s, L\rangle + \langle \eta_v M^s, L\rangle \\ &\geq \frac{1}{k}\langle \mathbb{1}, L\rangle - \|\eta_v\|_1 \|L\|_\infty \\ &\geq \frac{1}{k}\langle \mathbb{1}, L\rangle - \frac{kc\|L\|_\infty}{\sqrt{t}} \\ &> \frac{1}{2k}\langle \mathbb{1}, L\rangle,\end{aligned}$$

where the first inequality follows from Lemma 13. Let $d_2 = e^{\frac{1}{2k}\langle \mathbb{1}, L\rangle}$ and note that this implies that $v \in \Sigma_{\frac{1}{2k}\langle \mathbb{1}, L\rangle}$ and so by Theorem 9 there is a constant s such that with probability at least $1 - se^{-(\frac{1}{12}-o(1))d_2^t}$ the vertex v is connected to $\mathcal{S}_\epsilon$ by a path of length at most s. Observing that a constant fraction of the vertices have the desired signature by Chernoff bounds completes the proof. □

A slight modification of this argument gives part (3) of the main theorem.

Theorem 15. *Let G be a t^{th}-order stochastic Kronecker graph generated by a matrix $P \in [0,1]^{k\times k}$ such that W is connected and non-bipartite. If $\prod_i c_i = 1$ such that the c_i's are not all equal, then there are constants $s, d > 1$, depending only on P, such that for sufficiently large t, G has a giant component with probability at least $1 - e^{-\Theta\left(d^t\right)}$.*

Proof. Since all the c_i's are distinct, we have that $\sum_i c_i \ln(c_i) > 0$ by Lemma 12. Fix

$$0 < \epsilon = \frac{\sum_i c_i \ln(c_i)}{2\sum_i c_i \ln(c_i) - 2\ln(c_1)\operatorname{Vol}(W)} < \frac{\sum_i c_i \ln(c_i)}{\sum_i c_i \ln(c_i) - \ln(c_1)\operatorname{Vol}(W)}.$$

Again we have by Theorem 8 there is some constant $d_1 > 1$ such that $\mathcal{S}_\epsilon$ is connected with probability at least $1 - e^{-\Theta\left(d_1^t\right)}$.

Let $s = \left\lceil \frac{1}{\lambda} \ln\left(\frac{2\operatorname{Vol}(W)}{\epsilon c_1}\right)\right\rceil$ and note that by Theorem 3, $\frac{1}{k}\mathbb{1}M^j \in \mathcal{S}_{\epsilon/2}$ for all $j \geq s$. Thus $\left\langle \frac{1}{k}\mathbb{1}M^j, L\right\rangle \geq \epsilon c_1 + \sum_i (1-\epsilon)\pi \ln(c_i)$ for all $j \geq s$. Since s is a fixed constant, this implies that there is some $\nu > 0$ such that for all $j \geq 1$, we have $\left\langle \frac{1}{k}\mathbb{1}M^j, L\right\rangle \geq \nu$.

Let c be a constant to be fixed later. We notice that for t sufficiently large all vertices v such that $\left\|\sigma(v) - \frac{1}{k}\mathbb{1}M\right\| \leq \frac{c}{\sqrt{t}}$ are contained in $\Sigma_{\nu/2}$. Thus by Theorem 9 these vertices are connected to $\mathcal{S}_{\epsilon/2}$ with probability at least $1 - e^{-\Theta\left(d_2^t\right)}$ where $d_2 = e^{-\nu/2}$.

At this point it suffices to show that a constant fraction of the vertices in G are adjacent to $\Sigma_{\nu/2}$. To this end, consider the vertices $v \in V'$ such that $\left|\sigma_j(v) - \frac{1}{k}\right| \leq \frac{1}{k\sqrt{t}}$ for $1 \leq j < k$ and $\left|\sigma_k(v) - \frac{1}{k}\right| \geq \frac{-\ln(c_1)}{\ln(c_k)\sqrt{t}}$. By Chernoff bounds and Observation 7, we have that a constant fraction of the vertices of G are in V'. Furthermore, for every vertex $v \in V'$, $\mathbb{E}[\deg(v)] \geq 1$. Now by part (1) of Lemma 4, for all $v \in V'$,

$$\sum_{\|\sigma(u)-\sigma(v)M\| \leq \frac{1}{\sqrt{2t}}} \mathbb{P}(u \sim v) \geq (1 - e^{-1})\mathbb{E}[\deg(v)] \geq 1 - e^{-1}.$$

Thus, any fixed vertex in $v \in V'$ has a neighbor u such that $\|\sigma(u) - \sigma(v)M\| \leq \frac{1}{\sqrt{2t}}$ with probability at least $e^{-2(1-e^{-1})}$. Taking $c \geq \frac{1}{\sqrt{2}} + \max\left\{\frac{1}{k}, \frac{-\ln(c_1)}{\ln(c_k)}\right\}$ and applying Chernoff bounds completes the proof. □

6. Connectivity

Finally, we turn to the connectivity of G. We note that part (8) of the main theorem follows immediately from Theorem 2 by observing that the minimum degree in $W^{\otimes t}$ is exponential in t and exploiting the spectral properties of the Kronecker product. However, in keeping with the theme of this paper we provide an alternative proof which exploits the Markov chain structure.

Theorem 16. *Let G be a t^{th}-order stochastic Kronecker graph generated by a matrix $P \in [0,1]^{k\times k}$ such that W is connected and non-bipartite. If $1 < c_1 \leq \ldots \leq c_k$, then there is some constant $d > 1$, depending only on P such that G is connected with probability at least $1 - e^{-\Theta\left(d^t\right)}$.*

Proof. We first note that as $c_1 > 1$, $\ln(c_1) > 0$ and thus for any signature σ, $\langle \sigma, L\rangle \geq \ln(c_1) > 0$. Thus every vertex is in $\Sigma_{\ln(c_1)}$ and hence by Theorem 9 for every $\epsilon > 0$, every vertex is connected to $\mathcal{S}_\epsilon$ by a path of constant length with probability at least $1 - ne^{-c_1^{(1-o(1))t}}$. Thus it suffices to show that there is some $\epsilon > 0$ such that $\mathcal{S}_\epsilon$ is connected. But as $c_i > 1$ for all i, this implies that $\sum_i c_i \ln(c_i) > 0$ and thus by Theorem 8 there is some constant $\hat{d} > 1$, depending only on P, such that $\mathcal{S}_\epsilon$ is connected with probability at least $1 - e^{-\Theta\left(\hat{d}^t\right)}$. □

The following two theorems address the case that $c_1 = 1$. We note that we will always have a giant component in this case, unless $c_1 = c_2 = \cdots = c_k = 1$. However, the connectivity no longer depends entirely on the degrees in the graph, but is determined based on how the weight is distributed among the vertices. In particular, the backbone graph will determine the behavior.

Theorem 17. *Let G be a t^{th}-order stochastic Kronecker graph generated by $P \in [0,1]^{k\times k}$ with column sums $1 = c_1 \leq \cdots \leq c_k$. If W is connected and non-bipartite and the backbone graph B has a vertex of degree zero, then there is a constant $p \in (0,1)$ such that with probability at least $1 - p^t$ the graph G has at least $\frac{1}{2}t^{(1-o(1))\ln\ln(t)}$ isolated vertices.*

Proof. Fix a vertex v such that for all vertices u, $\mathbb{P}(u \sim v) \leq \frac{1}{2}$. We note that in this case we have

$$\begin{aligned}
\ln\left(\mathbb{P}(\deg(v) = 0)\right) &= \ln\left(\prod_u 1 - \mathbb{P}(u \sim v)\right) \\
&= \sum_u \ln\left(1 - \mathbb{P}(u \sim v)\right) \\
&\geq -\sum_u \frac{\mathbb{P}(u \sim v)}{1 - \mathbb{P}(u \sim v)} \\
&\geq -\sum_u 2\mathbb{P}(u \sim v) \\
&= -2\mathbb{E}[\deg v]\,,
\end{aligned}$$

where the last inequality comes from the upper bound on $\mathbb{P}(u \sim v)$. Thus we have that $\mathbb{P}(\deg(v) = 0) \geq e^{-2\mathbb{E}[\deg(v)]}$. Thus it suffices to find a large collection of vertices in G whose degrees are independent and where $\mathbb{E}[\deg(v)]$ is small.

To that end suppose that there is some i such that $p_{1i} = 1$, that is, the degree of vertex 1 in the B is not zero. Thus there is some $j \neq 1, i$ such that j has degree zero in B. Now let $S_{t_j}^{(j)}$ be the set of vertices in G whose signature σ has $\sigma_j = \frac{t_j}{t}$, $\sigma_1 = 1 - \frac{t_j}{t}$, and $\sigma_i = 0$ for $i \neq 0, j$. Since $c_1 = 1$ and $p_{1i} = 1$, we know that $p_{1j} = 0$ and thus the degrees of all vertices in $S_{t_j}^{(j)}$ are independent. We note that there is a choice of constant c such that if $t_j = c\ln\ln(t)$ then the expected number of isolated vertices in $S_{t_j}^{(j)}$ is $t^{(1-o(1))\ln\ln(t)}$, and thus by Chernoff bounds with probability at least $1 - e^{-\frac{t^{(1-o(1))\ln\ln(t)}}{6}}$ there are at least $\frac{1}{2}t^{(1-o(1))\ln\ln(t)}$ isolated vertices in G.

Now suppose that the degree of 1 in B is zero. Choice some index $j \neq 1$ arbitrarily and consider the set $S_{t_j}^{(j)}$ as above. As j is arbitrary there may be some edges between vertices of $S_{t_j}^{(j)}$. Thus we note that when

$2t_j \leq t$, we have

$$\begin{aligned}\mathbb{E}\Big[e(S_{t_j}^{(j)}, S_{t_j}^{(j)})\Big] &= \sum_{u \in S_{t_j}^{(j)}} \sum_{v \in S_{t_j}^{(j)}} \mathbb{P}(u \sim v) \\ &= 2\binom{t}{t_j} \sum_{i=0}^{t_j} \binom{t_j}{i} p_{jj}^{t_j - i} p_{j1}^{i} p_{1j}^{i} p_{11}^{t-t_j-i} \\ &= 2\binom{t}{t_j} p_{11}^{t-2t_j} \left(p_{11}p_{jj} + p_{1j}^2\right)^{t_j}.\end{aligned}$$

In particular, there is a constant c' such that $\mathbb{E}\Big[e(S_{t_j}^{(j)}, S_{t_j}^{(j)})\Big] \leq (c't)^{t_j}\, p_{11}^t$. As $p_{11} < 1$, this implies that the probability of an edge in $S_{t_j}^{(j)}$ is exponentially small provided $t_j \in o\Big(\frac{t}{\ln(t)}\Big)$. Thus, again choosing $t_j = c \ln\ln(t)$ and conditioning on $e(S_{t_j}^{(j)}, S_{t_j}^{(j)}) = 0$ gives the desired result. □

A slight simplification of this result gives part (5) of Theorem 1.

Theorem 18. *Let G be a t^{th}-order stochastic Kronecker graph generated by a matrix $P \in [0,1]^{k \times k}$ such that W is connected and non-bipartite. If $1 = c_1 \leq \ldots \leq c_k$ and the backbone graph B has no vertices of degree zero, then there is a constant $d > 1$ such that G is connected with probability at least $1 - e^{-\Theta\left(d^t\right)}$.*

Proof. First we note that $c_k > 1$ as otherwise the only edges present in the W are those present in the backbone graph, and in particular, W is a perfect matching contradicting the non-bipartiteness. Thus we have that $\sum_i c_i \ln(c_i) > 0$ and thus by Theorem 8 there is some $\epsilon > 0$ and $d' > 1$ such that $S_{2\epsilon}$ is connected with probability at least $1 - e^{-\Theta\left(d'^t\right)}$.

Now in a similar manner as the proof of Theorem 9 it suffices to show that with high probability that from every vertex $v = v_0$ there is a sequence $v_0, v_1, \ldots, v_s$ such that $v_i \sim v_{i+1}$ and $v_s \in \mathcal{S}_\epsilon \subset S_{2\epsilon}$. However, letting $s = \left\lceil \frac{1}{\lambda} \ln\left(\frac{\mathrm{Vol}(W)}{2\epsilon}\right)\right\rceil$ and imposing the additional condition that $\left\|\sigma(v_i)M - \sigma(v_{i+1})\right\|_\infty \leq \frac{\epsilon}{sk\|L\|_\infty}$, gives that $v_s \in \mathcal{S}_\epsilon$ by Theorem 3 and the Markov chain viewpoint.

To that end fix an arbitrary vertex v and consider the behavior of $Z^{(v)}$ from the point of view of the product distribution $\left(X^{(1)}\right)^{t_1} \times \cdots \times \left(X^{(1)}\right)^{t_k}$ where t_i is the number of i's the label for v. Notice that for those indicies i where $c_i = 1$, $X^{(i)}$ is the identity distribution. Furthermore, these coordinates perfectly respect the action of the Markov chain given by M. Suppose then that $t_j + \cdots + t_k \leq \frac{\epsilon}{sk\|L\|_\infty} t$, then any neighbor u of v in $B^{\otimes t}$ satisfies that $\left\|\sigma(v)M - \sigma(u)\right\|_\infty \leq \frac{\epsilon}{sk\|L\|_\infty}$. Otherwise, $\mathbb{E}[\deg(v)] \geq c_j^{\frac{\epsilon}{sk\|L\|_\infty} t}$ and by Lemma 4, there is a constant c such that

$$\sum_{\left\|\sigma(v)M - \sigma(u)\right\|_\infty} \mathbb{P}(u \sim v) \geq c c_j^{\frac{\epsilon}{sk\|L\|_\infty} t}.$$

Applying Chernoff bounds to assure the existence of such a vertex completes the proof. □

7. Concluding Remarks

We note that in principle these techniques can be extended to analyze the emergence of connectivity and the giant component in generalizations of the stochastic Kronecker graph, such as the multiplicative attribute graph [13]. In fact, based on the work in [20], it is likely that similar transition points will hold. That is, the multiplicative attribute graph will have a giant component when the median expected degree is 1 and become connected when the probability of an isolated vertex goes to zero.

Perhaps a more interesting direction would to resolve the size of the largest component in the case when $c_1 = c_2 = \cdots = c_k = 1$. By letting $P = \frac{1}{k}\mathbb{1}\mathbb{1}^T$ we see that this regime includes the Erdős-Rényi graph $\mathcal{G}\left(k^t, \frac{1}{k^t}\right)$ at criticality. Thus it seems likely that in order to understand the size of the largest component of the stochastic Kronecker graph when $c_1 = c_2 = \cdots = c_k = 1$ it will require a deeper understanding of why the branching process for $\mathcal{G}(n, \frac{1}{n})$ terminates with a largest component of size $\Theta\left(n^{2/3}\right)$ [2].

As a possible intermediate stage, consider a d-regular, connected, non-bipartite graph H on k vertices and let P be $\frac{1}{d}$ times the adjacency matrix of H. What is the size of the largest component in the t^{th}-order stochastic Kronecker graph generated by P? From a natural coupling with $\mathcal{G}\left(d^t, \frac{1}{d^t}\right)$ it is clear that it should be at least $\Omega\left(d^{2t/3}\right)$. On the other hand, since the degree of every vertex is still asymptotically Poisson with parameter 1, the branching process point of view would indicate that the size of the largest component should be $\Theta\left(k^{2t/3}\right)$. However, we note that if H is the d-regular graph formed by two copies of K_{d-1} joined by a perfect matching, then $H^{\otimes t}$ consists of 2^t copies of $K_{(d-1)^t}$ with relatively few edges between them. Furthermore, as the expected degree within each of these copies of $K_{(d-1)^t}$ is $\left(\frac{d-1}{d}\right)^t \in o(1)$, the largest component in each of these components is $\mathcal{O}(t)$, seemingly indicating that the overall size of the largest component is relatively small. Thus, it seems likely that any resolution of the case where $c_1 = c_2 = \cdots = c_k$ will necessitate a deeper understanding of the branching process at criticalility, and specifically, how the branching process interacts with the underlying network of potential edges.

References

[1] Béla Bollobás, *The diameter of random graphs*, Trans. Amer. Math. Soc., 267 (1981), pp. 41–52.

[2] ———, *The evolution of random graphs*, Trans. Amer. Math. Soc., 286 (1984), pp. 257–274.

[3] Béla Bollobás and Oliver Riordan, *Asymptotic normality of the size of the giant component via a random walk*, J. Combin. Theory Ser. B, 102 (2012), pp. 53–61.

[4] ———, *A simple branching process approach to the phase transition in $G_{n,p}$*, Electron. J. Combin., 19 (2012), pp. Paper 21, 8.

[5] Fan Chung, Paul Horn, and Linyuan Lu, *The giant component in a random subgraph of a given graph*, in Algorithms and models for the web-graph, vol. 5427 of Lecture Notes in Comput. Sci., Springer, Berlin, 2009, pp. 38–49.

[6] ———, *Percolation in general graphs*, Internet Math., 6 (2009), pp. 331–347 (2010).

[7] Fan Chung and Linyuan Lu, *Connected components in random graphs with given expected degree sequences*, Ann. Comb., 6 (2002), pp. 125–145.

[8] Fan R. K. Chung, *Spectral graph theory*, vol. 92 of CBMS Regional Conference Series in Mathematics, Published for the Conference Board of the Mathematical Sciences, Washington, DC, 1997.

[9] Jian Ding, Jeong Han Kim, Eyal Lubetzky, and Yuval Peres, *Anatomy of a young giant component in the random graph*, Random Structures Algorithms, 39 (2011), pp. 139–178.

[10] P. Erdős and A. Rényi, *On the evolution of random graphs*, Magyar Tud. Akad. Mat. Kutató Int. Közl., 5 (1960), pp. 17–61.

[11] Alan Frieze, Michael Krivelevich, and Ryan Martin, *The emergence of a giant component in random subgraphs of pseudo-random graphs*, Random Structures Algorithms, 24 (2004), pp. 42–50.

[12] Svante Janson and Joel Spencer, *Phase transitions for modified Erdős–Rényi processes*, Ark. Mat., 50 (2012), pp. 305–329.

[13] Myunghwan Kim and Jure Leskovec, *Multiplicative attribute graph model of real-world networks*, in 7^{th} Workshop on Algorithms and Models for the Web Graph, 2010. preprint, arXiv:1009.3499v3.

[14] Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, and Christos Faloutsos, *Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication*, in European Conference on Principles and Practice of Knowledge Discovery in Database, 2005.

[15] Jure Leskovec and Christos Faloutsos, *Scalable modeling of real graphs using kronecker multiplication*, in ICML '07: Proceedings of the 24th international conference on Machine learning, New York, NY, USA, 2007, ACM, pp. 497–504.

[16] Tomasz Łuczak, *Component behavior near the critical point of the random graph process*, Random Structures Algorithms, 1 (1990), pp. 287–310.

[17] Mohammad Mahdian and Ying Xu, *Stochastic Kronecker graphs*, in Algorithms and models for the web-graph, vol. 4863 of Lecture Notes in Comput. Sci., Springer, Berlin, 2007, pp. 179–186.

[18] Mary Radcliffe and Fan Chung, *On the spectra of general random graphs*. preprint.

[19] Mary Radcliffe and Paul Horn, *Giant components in kronecker graphs*, Random Structures & Algorithms, (2012).

[20] Mary Radcliffe and Stephen J. Young, *The spectra of multiplicative attribute graphs*. submitted, February 2012.

[21] Alistair Sinclair, *Algorithms for random generation and counting*, Progress in Theoretical Computer Science, Birkhäuser Boston Inc., Boston, MA, 1993. A Markov chain approach.

[22] Alistair Sinclair and Mark Jerrum, *Approximate counting, uniform generation and rapidly mixing Markov chains*, Inform. and Comput., 82 (1989), pp. 93–133.

[23] Joel Spencer, *The giant component: the golden anniversary*, Notices Amer. Math. Soc., 57 (2010), pp. 720–724.

Department of Mathematics, University of Washington, Seattle, WA 98195
E-mail address: radcliffe@math.washington.edu

Department of Mathematics, University of Louisville, Louisville, KY 40292
E-mail address: stephen.young@louisville.edu