On the Spectrum of Dense Random Multigraphs

A. Montanari, P. Syminelakis

October 19, 2014

Abstract

Spectral properties of random matrices are important in various fields from Physics, Mathematics to Signal Processing and Computer Science. The typical assumption is that of independent entries (up to a symmetry class) under which the Four Moments Theorem [Tao-Vu] shows universality of Wigner Semicircle Law. In this work, we consider a model of random graphs (matrices) where a number of small (finite) subgraphs are inserted in an initially empty graph. We show that despite the fact that our model exhibits strong local dependencies the limiting spectral distribution is again the Wigner Law. Our result along with the universality of Wigner law argues against using Spectral Statistics for distinguishing network models.

1 Introduction

In this report we consider a simple model of dense random multigraphs and investigate the limiting spectral distribution (LSD) of the associated adjacency matrix. In particular, we consider the model where $m$ finite graphs are inserted uniformly at random, increasing the weight of the corresponding edges each time. The motivation for investigating this kind of model lies in probing the range of possible limiting distributions of the eigenvalues of random graphs.

The spectrum of random graphs has been extensively investigated [2]. There are three flavors of results: (i) analytical: based on methods of local weak convergence, (ii) combinatorial: based on the moment method and enumeration of paths and (iii) results based on matrix concentration inequalities. The results obtained by these methods characterize the LSD in cases where the graphs are locally treelike (random regular, random sparse) [4, 10, 8, 3] or have independent edges [7]. There are also some results in cases when the edges are exchangable [6], and when they are uncorrelated using martingale conditions [9, 1].

This work differentiates with respect to the existing literature by relaxing the independence and local tree likeness assumptions. In particular, we are interested in finding whether short range correlations between edges are enough to alter the asymptotic behaviour of the spectrum and whether the LSD is affected by the specific graph considered.

We consider the following model of a random graph $G_n$ on $n$ vertices. Let $\mu_n : \mathcal{F} \to [0,1]$ be a sequence of discrete probability measures over a collection of $q$ labeled finite graphs $\mathcal{F} = \{F_1, \ldots, F_q\}$ assigning probabilities $\mu_n(F_i) = p_i(n)$ to each such graph. The graph is constructed by the following process, for each time $\ell = 1, \ldots, m(n)$:

1. Sample a labeled graph $F^{(\ell)}$ from $\mathcal{F}$ according to $\mu_n$: $P(F^{(\ell)} = F_i) = \mu_n(F_i) = p_i(n)$.

2. Let $v = |V(F^{(\ell)})|$ and $r = |E(F^{(\ell)})|$. Sample a set of $v$ vertices of $G_n$ uniformly at random and assign labels by considering a random permutation $\pi$ of the vertices.
3. For each edge \((i,j) \in E(F^{(\ell)})\) add the edge \((\pi^{-1}(i), \pi^{-1}(j))\) in \(G_n\), where \(\pi^{-1}(i)\) denotes the vertex of \(G_n\) that was assigned label \(i\).

In order to study such graphs we consider the corresponding symmetric adjacency matrix \(A_n \in \mathbb{R}^{n \times n}\) of \(G_n\). In particular let \(A_n^{(\ell)}\) be the zero mean adjacency matrix corresponding to the edges inserted in \(G_n\) at time \(\ell\):

\[
A_n^{(\ell)} = \begin{cases} 
\alpha, & \text{w.p. } \rho \\
-\alpha \frac{1}{\rho}, & \text{w.p. } 1 - \rho 
\end{cases}
\]

where \(\rho(n)\) is the probability of the edge \((i,j)\) being selected as part of the graph inserted at time \(\ell\), \(\alpha\) a regularizing parameter. The adjacency matrix of the final multigraph is then \(A_n = A_n^{(1)} + \ldots + A_n^{(m)}\). We consider that \(m(n) = dna(n)\), where \(a(n) \to \infty\) with \(n\), \(\alpha(n) = 1/\sqrt{a(n)}\). Under this regime we prove the following theorem:

**Theorem 1** The empirical spectral distribution of the adjacency matrix \(A_n = A_n^{(1)} + \ldots + A_n^{(m)}\) converges to the Wigner semicircle law supported on \([-R, R]\) with \(R\) given by:

\[
R(\mu, F, d) = \sqrt{8d\mu(r(F))}
\]

where \(r(F) = |E(F)|\) denotes the number of edges of a graph in \(F\), \(\mu_n \to \mu\) and \(m(n) = dna(n)\), with \(a(n) \to \infty\).

We prove our theorem by standard arguments using the Moment Method. The core of the proof lies with Lemma 1, which essentially asserts that only dependencies due to symmetry of the adjacency matrix matter. We can draw two conclusions from the theorem. The first is that the empirical spectral distribution depends only on the mean number of edges inserted and not on the particular collection of graphs \(F\). Secondly, that finite range correlations combined with uniform sampling are not sufficient to induce interesting artifacts in the spectral distribution. An intuitive explanation of this fact can be derived by considering a factor graph representation of our graph, where each element \((i,j)\) of the matrix \(A_n\) is a node and each small subgraph \(F^{(\ell)}\) is a constraint node. We add an edge between element \((i,j)\) and constraint \(\ell\), if the edge \((i,j)\) was selected as part of the graph \(F^{(\ell)}\) inserted at time \(\ell\). The resulting factor graph is approximately Poisson with mean of
the order \(a(n)/n\), which means that even for \(a(n) = n\), i.e. \(m(n) = n^2\), the correlations between the elements of the matrix are very weak.

An interesting question that arises is, whether introducing the freedom of choice can alter our results and if so in what extent. In particular, we could consider “Ramsey Games” in this ensemble, where instead of sampling one copy of the graph \(F^{(\ell)}\) at each time, we sample \(k\) copies (different collection of vertices to insert edges) and we have freedom to keep one of the \(k\). How deformed can the spectrum be?

2 Proof of Theorem 1

In order to prove our theorem we use the Method of Moments. We first use independence of the edges added at different times \(\ell\) to reduce the calculation of moments to a calculation involving partitions of edges in blocks along a path. Then, we show that only partitions with blocks of size two matter and that from those partitions only non-crossing partitions of a specific form have non-negligible contribution. Finally, we use these results to specify the limits of moments of \(A_n\) and showing that they coincide with the moments of the Wigner Semicircle law.

We begin by calculating the probability \(\rho(F) = \mathbb{P}(A^{(\ell)}_{ij} = \alpha | F^{(\ell)} = F)\) that a particular edge \((i, j)\) is selected as part of the graph \(F\) inserted at time \(\ell\). Let \(v = |V(F)|\) and \(r(F) = |E(F)|\) be the number of vertices and edges of the graph \(F\), then:

\[
\rho(F) = \frac{\binom{n}{v - 2} 2r(F)(v - 2)!}{\binom{n}{v} v!} = \frac{2r}{(n - v + 1)(n - v + 2)} \approx 2rF n^{-2}
\]

Since, there are \(r\) labeled edges and we can label the vertices of our specific edge \((i, j)\) in two ways for each such edge. The rest \((v - 2)\) vertices selected in the set can be labeled arbitrarily. We note that the elements of \(A^{(\ell)}\) are not independent and have marginals given by:

\[
A^{(\ell)}_{ij} = \begin{cases} 
\alpha, & \text{w.p. } \rho \\
-\alpha \rho, & \text{w.p. } 1 - \rho
\end{cases}
\]

for all \(i \neq j\), since \(1 - \rho \simeq 1\) and \(A_{ii} = 0\) for all \(i\). Recall that \(\alpha = 1/\sqrt{a(n)}\) and \(m = dna(n)\) so that the matrix satisfies \(\mathbb{E}A = 0\) and:

\[
\frac{1}{n} \mathbb{E}trA^2 = \sum_{j=1}^{n} \mathbb{E}(A_{ij}A_{ji}) = \sum_{j=1}^{n} \sum_{\ell=1}^{m(n)} \mathbb{E}(A^{(\ell)}_{ij}A^{(\ell)}_{ji})
\]

\[
= \sum_{j=1}^{n} \sum_{\ell=1}^{m(n)} p \sum_{s=1}^{p} \mathbb{E}(A^{(\ell)}_{ij}A^{(\ell)}_{ji} | F^{(\ell)} = F_s)\mu_n(F_s)
\]

\[
= \sum_{s=1}^{p} \mu_n(F_s) \left( \sum_{j=1}^{n} \sum_{\ell=1}^{m(n)} \mathbb{E}(A^{(\ell)}_{ij}A^{(\ell)}_{ji} | F^{(\ell)} = F_s) \right)
\]

\[
\simeq \sum_{s=1}^{p} \mu_n(F_s) \left[ nm(n)\alpha^2 \rho(F_s) \right] = 2d\mu_n(r(F))
\]
where we have used that the matrices $A^{(\ell)}$ are independent and that $\mathbb{E}(A_{ij}^{(\ell)}) = 0$. Having defined the model and calculated these simple quantities we turn next to the issue of calculating the limits:

$$
\beta_k = \lim_{n \to \infty} \beta_{k,n} = \lim_{n \to \infty} \frac{1}{n} \mathbf{tr} A^k
$$

(4)

If the limits $\beta_k$ satisfy Carleman’s condition, we can use the Moment Convergence Theorem (see [2]) and identify the LSD as the unique distribution with moments $\beta_k$. We start with some definitions [11]:

**Definition 1 (Partitions).**
- We call $\pi = \{V_1, \ldots, V_b\}$ a partition of a set $S = (1, \ldots, n)$ if and only if $V_i (1 \leq i \leq b)$ are pairwise disjoint, non-void tuples such that $V_1 \cup \ldots \cup V_b = S$.
- We call the tuples $V_1, \ldots, V_b$ the blocks of the partition $\pi$.
- Given two elements $p$ and $q$, we write $p \pi \sim q$, if $p$ and $q$ belong in the same block of the partition.

**Definition 2 (Non-crossing Partitions)** A partition is called non-crossing if there are no indices $p_1 < q_1 < p_2 < q_2$ such that $p_1 \pi \sim p_2 \pi \sim q_1 \pi \sim q_2$. Another recursive definition is that a partition is non-crossing if at least one block $V$ is an interval and $\pi \setminus V$ is non-crossing. Let $NC(1, \ldots, n)$ be the set of non-crossing partitions of $n$ labeled elements, then:

$$
\pi \in NC(1, \ldots, n) \iff \exists V = (k, \ldots, k+p) \text{ such that } \pi \setminus V \in NC(1, \ldots, k-1, k+p+1, \ldots, n)
$$

(5)

Using this recursive definition it is easy to show via induction that if $D_i$ is the number of non-crossing partitions with $i$ elements then:

$$
D_n = \sum_{i=1}^{n} D_{i-1} D_{n-i}
$$

(6)

which is the recursion for the Catalan numbers $C_n$. Since, also $D_1 = C_1 = 1$, we get that $D_n = C_n$.

**Definition 3 (Consistent Sequences)** A sequence of edges $P_1, \ldots, P_k$ is called consistent if for all edges $P_\ell = (s_\ell, t_\ell)$ we have that:

$$
s_i = t_{i-1}, \ i = 2, \ldots, k \quad \text{and} \quad t_k = s_1
$$

(7)

Using the above concepts, next we reduce the calculation of moments in a calculation involving partitions of edges forming a consistent sequence. We have that:

$$
\mathbf{tr} A^k = \sum_{i=1}^{n} (A^k)_{ii} = \sum_{i=1}^{n} \left( \sum_{P_1, \ldots, P_k} A_{P_1} \ldots A_{P_k} \right)_{ii}
$$

(8)
where $P_1, \ldots, P_k$ is a consistent sequence. Using the fact that for each edge $P_e$, $A_{P_e} = \sum_{\ell=1}^m A_{P_e}^{(\ell)}$, we get:

$$
\text{tr} A^k = \sum_{i=1}^n \sum_{P_1^{(i)}, \ldots, P_k^{(i)}} \left( \sum_{\ell_1=1}^m A_{P_{\ell_1}}^{(\ell_1)} \cdots \sum_{\ell_k=1}^m A_{P_{\ell_k}}^{(\ell_k)} \right) 
$$

$$
= \sum_{i=1}^n \sum_{P_1^{(i)}, \ldots, P_k^{(i)}} \sum_{\ell_1, \ldots, \ell_k} \left( A_{P_{\ell_1}}^{(\ell_1)} \cdots A_{P_{\ell_k}}^{(\ell_k)} \right)
$$

Now, due to symmetry (permutation invariance of the vertices), we get that:

$$
\beta_{k,n} = \frac{1}{n} \mathbb{E} \text{tr} A^k = \sum_{P_1, \ldots, P_k} \sum_{\ell_1, \ldots, \ell_k} \mathbb{E} \left[ A_{P_{\ell_1}}^{(\ell_1)} \cdots A_{P_{\ell_k}}^{(\ell_k)} \right] 
$$

(9)

where we have fixed without loss of generality the first edge to be of the form $P_1 = (1, i)$ and the last edge $P_k = (j, 1)$. Now, each sequence of indices $\ell_1, \ldots, \ell_k$ defines a partition over the edges $P_1, \ldots, P_k$, where edges $P_x, P_y$ such that $\ell_x = \ell_y$ belong in the same partition. Furthermore, depending on the actual value of the index $\ell$ we can assign a “color” to the specific block of the partition. Edges belonging to different partitions are independent, since the matrix $A$ is a sum of independent matrices each corresponding to an insertion of the small graph $F^{(\ell)}$. There are $m$ “colors” and hence:

$$
\beta_{k,n} = \sum_{\pi \in P(k)} \binom{m}{b_{\pi}} b_{\pi}! \sum_{P_1, \ldots, P_k} \prod_{\ell=1}^{L_{\pi}} \mathbb{E} \left[ A_{P_{\ell_1}}^{(\ell_1)} \cdots A_{P_{\ell_k}}^{(\ell_k)} \right] 
$$

(10)

where $b_{\pi} \leq k$ is the number of blocks of the partition $\pi$. Since $m(n) \to \infty$ with $n$ and $b_{\pi}$ is constant, we get that:

$$
\beta_{k,n} = \sum_{\pi \in P(k)} \sum_{P_1, \ldots, P_k} \prod_{\ell=1}^{L_{\pi}} \mathbb{E} \left[ m \cdot \prod_{i=1}^{L_{\pi}} A_{P_{\ell_i}} \right] 
$$

(11)

In order to count the number of consistent sequences $P_1, \ldots, P_k$, we consider that always the first index $s_1$ of an edge $P_1 = (s_1, t_1)$ is fixed (due to consistency) by the second index $t_{i-1}$ of the previous edge, or is 1 if $i = 1$. So, in general we can only pick the indices $t_i$ for $1 \leq i \leq k - 1$, since the last index $t_k$ is fixed by consistency to be 1. Therefore, there is a bijection between consistent sequences and numbers $t_1, \ldots, t_k$. Moreover, accounting for the number of sequences in this way, decouples
In order to prove the lemma we need only to calculate the probability that \( S \), \( \pi \) is a partition of \( E \), the graph \( F \) inserted at time \( \ell \). Observe that these events are disjoint for different sets \( S \) and we can write:

\[
E \left[ m \cdot \prod_{i=1}^{\ell} A_{P_i} \right] = m a^L \sum_{S \in \mathcal{P}(V_i)} \rho^{L-|S|} P(B_\ell(S))
\]

**Lemma 1 (Bounding Block Weight)** Let \( S^*_\ell \subseteq V_\ell \) be the set of maximum cardinality such that \( S^*_\ell \subseteq E(F) \), and \( v_\ell = V[S^*_\ell] \) be the number of different vertices that appear in \( S^*_\ell \), then:

\[
E \left[ m \cdot \prod_{i=1}^{\ell} A_{P_i} \right] \leq C_{L,F} a(n)^{1-L/2} n^{1-v_\ell-2(L-|S^*_\ell|)}
\]

where \( C_{L,F} = d(4r(F))^L v(F)! \) is a finite constant that does not depend on \( n \).

**Proof.** In order to prove the lemma we need only to calculate the probability \( P(B_\ell(S)) \) for a given \( S \subseteq V_\ell \). Let \( v_S \) be the number of vertices induced by \( S \). The probability then is upper bounded by:

\[
P(B_\ell(S)) \leq \mathbb{I}_{S \subseteq E(F)} \frac{\binom{n}{v_S} v_S!}{(n-v_S)!} \frac{(n-v_S)!}{v!} \leq n^{-v_S} v_S! \mathbb{I}_{S \subseteq E(F)} \leq v! n^{-v_S} \mathbb{I}_{S \subseteq E(F)}
\]

Using the last inequality we have that:

\[
\rho^{L-|S|} P(B_\ell(S)) \leq (2r(F))^L v! n^{2|S|-v_S-2L} \mathbb{I}_{S \subseteq E(F)}
\]

Hence, the maximizing \( S \) is the one with maximum cardinality for which \( S \subseteq E(F) \), since \( v_S \) can increase at most by two for every edge that we include in \( S \). Putting it all together:

\[
E \left[ m \cdot \prod_{i=1}^{\ell} A_{P_i} \right] = m a^L \sum_{S \in \mathcal{P}(V_i)} \rho^{L-|S|} P(B_\ell(S)) \leq (4r(F))^L v! da(n)^{1-L/2} n^{1-v_\ell-2(L-|S^*_\ell|)}
\]

where we used the fact that \( m = dna(n), \rho \simeq 2r(F)n^{-2} \) and \( \alpha = a(n)^{-1/2} \). \( \square \)

**Lemma 2 (Bounding Partition Weight)** Given a partition \( \pi \) with \( b_\pi \) blocks of size \( L_\ell > 1 \) for \( \ell = 1, \ldots, b_\pi \), the total contribution is bounded by:

\[
\sum_{P_1, \ldots, P_k} \prod_{\ell=1}^{b_\pi} \mathbb{E} \left[ m \cdot \prod_{i=1}^{L_\ell} A_{P_i} \right] \leq C_\pi a(n)^{b_\pi - \frac{\sum_{\ell=1}^{b_\pi} L_\ell}{2}}
\]

the choice of indices between different edges and therefore blocks. We will use this counting scheme to bound the contribution of each block of the partition to the total. From now on, we assume that \( \mathcal{F} \) consists only of a single graph \( F \) and we show in the end how to treat the general case.

We need to deal only with partitions that have blocks of length greater than 1, since \( E(A_{ij}^{(\ell)}) = 0 \) for all \( \ell \) and \((i, j)\). Given a partition \( \pi \) and a consistent sequence of edges \( P_1, \ldots, P_k \), let \( L_\ell > 1 \) be the size of a block \( V_\ell \) of the partition. For every set \( S \subseteq V_\ell \) define the event:

\[
B_\ell(S) = \{ A_{P_\alpha} = \alpha, \forall s \in S \} \cap \{ A_{P_\alpha} = -\alpha \rho, \forall s \in V_\ell \setminus S \}
\]

which denotes the fact that from the edges in \( V_\ell \), only the edges in set \( S \) were selected as part of the graph \( F \) inserted at time \( \ell \).
Proof. First consider the case \( b_π = 1 \). We partition the admissible sequences according to the values of \( v_1 = V[S^*_1] \leq k \) (due to consistency) and \( w = |S^*_1| \leq k \). The total number of possible values for \((v_1, w)\) is at most \( k^2 \). In order to upper bound the number of admissible sequences of type \((v_1, w)\), we note that we can only select \( v_1 - 1 \) labels for the second indices of edges in the set \( S^*_1 \) (which depends on the actual sequence and not only the type) and at most \( L - w \) second indices of each edge in \( V_1 \setminus S^*_1 \). To see the first point, observe that one of the \( v_1 \) vertices must be the first vertex of the first edge, which is always fixed. Hence, using Lemma 1 we have that:

\[
\sum_{\{P_i\}_{i \in V_1}} \mathbb{E} \left[ m \cdot \prod_{t \in V_1} A_{P_t} \right] \leq \sum_{v_1, w} n^{v_1 - 1 + L - w} C_{L,F} a(n)^{1 - L/2} n^{1 - v_1 - 2(L - w)}
\]

where we used the fact that \(-L - w\) is always non-negative. This concludes the base case and suggests how to treat the general case as well. For the general case, we partition the admissible sequences in classes according to the values of \((v_1, \ldots, v_{b_π}, w_1, \ldots, w_{b_π})\) where each element is at most \( k \). A trivial upper bound on the number of such vectors is \( k^{2b_π} \). Thus, in a similar fashion with the case \( b_π = 1 \) we get that:

\[
\sum_{P_1, \ldots, P_k} \prod_{\ell = 1}^{b_π} \mathbb{E} \left[ m \cdot \prod_{i=1}^{L_{ij}} A_{P_{ki}} \right] \leq C_{k,F} k^{2b_π} a(n)^{b_π - \frac{\sum_{\ell=1}^{b_π} L_{ij}}{2}}
\]

so \( C_π = C_{k,F} k^{2b_π} \). The arguments carry on in the general case since we have decoupled the choice of indices for different blocks by first conditioning on the type of “block-maximizer” for the whole sequence and then considering that we only need to choose the second index for each edge. \( \square \)

Observing the previous proof, we note that the contribution of each block of the partition even when accounting for consistent sequences is bounded by a constant. This is easily seen by essentially the same arguments used in the proof of Lemma 2 for \( b_π = 1 \). Thus, if for one partition there is a block that has contribution tending to 0 with \( n \), then that partition is

**Corollary 1 (Only Symmetry Matters)** The contribution from partitions with any block length bigger than two is negligible. Furthermore, we have the following formula:

\[
\beta_{2k,n} \simeq \sum_{π \in PP(2k)} \sum_{j_1, \ldots, j_k} \prod_{i=1}^{k} \mathbb{E}[m A_{ij_1} A_{j_2 i_2}]
\]

where \( PP(2k) \) denotes the set of partitions of \( 2k \) elements in blocks of size 2.

**Proof.** Using Lemma 2, the fact that \( L_\ell > 1 \) for all \( \ell = 1, \ldots, b_π \) and that \( a(n) \to \infty \) with \( n \), we conclude that only pair partitions, where \( L_\ell = 2 \), have non-negligible contribution to \( \beta_{k,n} \). Since \( 2b_π - \sum_{\ell=1}^{b_π} L_\ell < 0 \) for any partition whenever there is a block of size greater than 2. Now, all blocks of the partition will have one of the three forms \( \mathbb{E}(A_{ij} A_{ik}), \mathbb{E}(A_{ij} A_{jk}), \mathbb{E}(A_{ij} A_{ji}) \). For a given pair partition \( π \), we partition the set of consistent sequences according to the number of elements that have each of the three forms. That is according to the vector \((z_1, z_2, z_3)\), where \( z_i \) is the number of blocks that have form \( i = 1, 2, 3 \). Now given a such a vector and a partition, we upper bound the contribution \( W_π(z_1, z_2, z_3) \) to the sum. Using the counting scheme for consistent sequences where we only select the second index we get that for any block:
• \( \sum_{j,\ell} E(mA_{ij}A_{\ell j}) = \Theta(n^2 \cdot na(n) \cdot a(n)) = \Theta(n^{-1}) \rightarrow 0 \)
• \( \sum_{j,\ell} E(mA_{ij}A_{\ell j}) = \Theta(n^2 \cdot na(n)) = \Theta(1) \)
• \( \sum_j E(mA_{ij}A_{ji}) = \Theta(n \cdot n^2 \cdot a^2) = \Theta(1) \)

Hence, the total contribution is:

\[
W_\pi(z_1, z_2, z_3) \leq 3^k \left( \frac{1}{n} \right)^I_{z_1 \lor z_2 > 0} C_\pi n^{-z_1}
\]

where \( 3^k \) is an upper bound on the number of ways to assign types to the \( k \) blocks of the partition. The term involving \( I_{z_1 \lor z_2 > 0} \) is due to the fact that in multiplying the contribution of the blocks, we must divide by \( n \) to correct for over-counting the ways to select the last index of the last edge, which is fixed by consistency. The last equation along with the fact that the number of such vectors \((z_1, z_2, z_3)\), where the indicator function is 1, is bounded by \( k^3 \), gives us that we only need to consider for a given pair partition sequences of the family \((0, 0, k)\), which gives us the statement of the lemma.

\( \square \)

**Lemma 3** Only non-crossing pair partitions (NCPP) have non-negligible contribution.

**Proof.** First we show that if the partition is non-crossing and all blocks are of the form \( \mathbb{E}(A_{ij}A_{ji}) \), the last index will always be equal to the first and, thus, we do not overcount. We use strong induction on \( b \) the number of blocks. For \( b = 1 \), it is trivial. Assume that it holds for \( b = k \), we show it for \( b = k + 1 \). Let \( 2 \leq i \leq 2(k + 1) \) be the index of the second edge of the first block of the partition, where we assume wlog that the the first edge belongs to the first block. We know that the first edge is of the form \((1, j)\) so \( P_i = (j, 1) \). By definition of a non-crossing partition we have that \( \pi = \{ A_{P_1}, A_{P_2} \} \cup \pi_{-i} \cup \pi_{i+} \), where \( \pi_{-i} \in NCPP(A_{P_2}, \ldots, A_{P_{i+1}}) \) and \( \pi_{i+} \in NCPP(A_{P_{i+1}}, \ldots, A_{P_{2(k+1)}}) \). Now, by consistency \( P_{i+1} \), the first edge of the non-crossing partition involving \( k - i + 1 < k + 1 \) blocks, is of the form \((1, \ell)\) and by strong induction we get that the last edge of the non-crossing pair partition \( \pi_{i+} \) with number of blocks at most \( k \) is of the form \((s, 1)\). Hence, we have showed that a non-crossing pair partition always ends with the index it starts.

Next, we show that if there is at least one crossing in the partition, then the partition it has negligible contribution. A crossing means that there are a indices \( p_1 < q_1 < p_2 < q_2 \) such that \( p_1 \sim p_2 \) and \( q_1 \sim q_2 \). Consider, minimal such index \( p_2 \) and a maximal such index \( q_1 \). This means that either \( p_2 = q_1 + 1 \) or the indices(edges) between \( q_1 \) and \( p_2 \), belong to a non crossing partition. Because otherwise there would be either a smaller index \( p'_2 \) or a larger index \( q'_1 \). Using the property that a consistent sequence belonging to a non-crossing partition, starts and ends with the same index, we get that for such indices where \( P_i = (s_i, t_i) \), we have:

\[
t_{q_1} = s_{p_2} = t_{p_1}
\]

where the second equality is due to the fact that we only need to consider blocks of the type \( \mathbb{E}(A_{ij}A_{ji}) \). Hence, each crossing results in at least one less degree of freedom for selecting a consistence sequence, since \( t_{q_1} = t_{p_1} \). So, to correct for overcounting, we need to divide the total contribution of such a partition with \( n \).
Using the last two lemmata we have the following equation up to negligible terms:

\[
\beta_{2k,n} = \sum_{\pi \in NCPP(2k)} \sum_{j_1, \ldots, j_k} \prod_{\ell=1}^{k} \mathbb{E}[m A_{i_\ell j_\ell} A_{j_\ell i_\ell}]
\]

At this point we return to the general case where instead of inserting a single graph \(F\), we sample each time a graph from \(\mathcal{F}\) according to a measure \(\mu_n\). Actually such generalization affects only the constants in lemmata 1, 2. Thus, the results concerning which partitions and sequences have negligible contribution still hold. However, in order to have exact results we need to work again in the general case:

\[
\beta_{2k,n} = \sum_{\pi \in NCPP(2k)} n^k \prod_{i=1}^{k} \left( \sum_{x=1}^{q} \mathbb{E}[m A_{i_\ell j_\ell} A_{j_\ell i_\ell} | F^{(\ell)} = F_x] \mu_n(F_x) \right)
\]

\[
= \sum_{\pi \in NCPP(2k)} \prod_{i=1}^{k} \left( \sum_{x=1}^{q} [nm^2 \rho(F_x)] \mu_n(F_x) \right)
\]

\[
= \sum_{\pi \in NCPP(2k)} \prod_{i=1}^{k} \left( \sum_{x=1}^{q} [nda(n)a(n)^{-1} 2r(F_x) n^{-2}] \mu_n(F_x) \right)
\]

\[
= \sum_{\pi \in NCPP(2k)} \prod_{i=1}^{k} \left( d2 \sum_{x=1}^{q} [r(F_x) \mu_n(F_x)] \right)
\]

\[
= \sum_{\pi \in NCPP(2k)} (2d \mu_n(r(F)))^k
\]

\[
= |NCPP(2k)| (2d \mu_n(r(F)))^k
\]

Thus to evaluate the moments in the limit we need \(\lim \mu_n(r(F)) = \mu(r(F))\) and the number of non-crossing pair partitions with \(2k\) elements. We can calculate the second number by using the fact that the number of non-crossing pair partitions \(Q_{2k}\) satisfies the recurrence relation:

\[
Q_{2k} = \sum_i Q_{2(i-1)} Q_{2(k-i)}
\]

Since, if the first block ends at index \(2i\) then the remaining \(2i - 2\) elements and the next \(2k - 2i\) elements must form non-crossing partitions. Thus, we have that \(Q_{2k} = C_k\), where \(C_k\) is the Catalan number. Finally:

\[
\beta_{2k,n} \to (2d \mu(r(F)))^k C_k = \left( \frac{R(\mu, \mathcal{F}, d)}{2} \right)^{2k} C_k
\]

with \(R(\mu, \mathcal{F}, d) = \sqrt{8 d \mu(r(F))}\). Therefore, we have identified the limit of the moments. By the moment convergence theorem we get that the LSD is the semicircle distribution with \(R(\mu, \mathcal{F}, d)\).

This concludes the proof of Theorem 1. 

\footnote{The index can only be an even number otherwise there must be a crossing.}
3 Application

We apply our results to study the spectral properties of the Subgraph Generation Model (SUGM) introduced by Chandrasekhar and Jackson [5]. They define the notion of a subgraph statistic $S^n_\ell(G)$ which is a count of how many instances of some particular (path-connected) subgraph on $v_\ell$ nodes appear in $G$. For instance $S^n_\ell(G)$ could be the number of triangles or the number of cliques of size 5 that appear in $G$. These statistics depending on the setting have a maximum value $\bar{S}_\ell$, e.g. $n^3_3$ for triangles in a uniform graph or $\binom{n}{5}$ for a clique of size 5. Further, for each statistic they consider a parameter $p^n_\ell$. Graphs in this model are constructed by adding in the graph each subgraph corresponding to a statistic $S^n_\ell$ independently with probability $p^n_\ell$, for each $\ell = 1, \ldots, k$. Thus, the statistics $S^n_\ell \sim \text{Bin}(\bar{S}_\ell, p^n_\ell)$ are binomial distributed random variables with mean $E(S^n_\ell) = \bar{S}_\ell p^n_\ell$. When, the parameters are such that $E(S^n_\ell) \to \infty$ with $n$, then by Chernoff bounds the statistic $S^n_\ell$ will be concentrated around its mean.

In order to treat this model in our framework, we need to consider an appropriate collection of finite graphs $\mathcal{F}$ and a sequence of measures $\mu_n : 2^{\mathcal{F}} \to [0,1]$. Hence, for each subgraph statistic $S^n_\ell$ we add the corresponding graph $F_\ell$ in the collection $\mathcal{F}$. Next, let $m_\ell(n) = E(S^n_\ell)$ be the expectation of a statistic according to the SUGM model. Let We define the discrete probability measure $\mu_n$ as:

$$
\mu_n(F_\ell) = \frac{m_\ell(n)}{\sum_i m_i(n)}
$$

and set $m(n) = \sum_i m_i(n)$, the total number of graphs that will be inserted. So, as long as $m(n) = \omega(n)$ is asymptotically larger than $n$, we can draw the following conclusions:

- The expected number of graph in both models is the same:

$$
E\left(\sum_i \mathbb{1}_{F_i}\right) = \mu_n(F) m(n) = m_\ell(n)
$$

- If $m_\ell(n) \to \infty$ both models will have identical subgraph statistics $S^n_\ell$ up to negligible terms due to concentration.

- The spectrum of the randomly generated graphs will be affected only by statistics such that $\mu_n(F_\ell) = \Omega(1)$.

- In the limit of large $n$ the spectrum will converge to the Wigner semicircle law.

To illustrate the applicability of our results, we use a dataset of real world networks out of 75 villages in India that Chandrasekhar and Jackson used in their simulations. We consider the collection $\mathcal{F} = \{K_3\}$ consists only of a triangle and we are going to compare the prediction about the maximum eigenvalue from our model with the actual one for each of the 75 villages. Since, we consider only a single graph we only need to know $m(n)$ the total number of triangles in the graph as a function of the number of vertices. Hence, for each village we counted the number of triangles and then fitted the curve $m(n) = dn^{1+\delta}$ using least squares. We found $d = 1.9283$ and $\delta = 0.246$. Our theoretical results predict that the spectrum of the normalized matrix $\tilde{A} = \frac{1}{\sqrt{m(n)}}(A - EA)$ will be supported on $[-R, R]$, where $R(d) = \sqrt{8 \cdot d \cdot 3} = 6.8029$, which is close to the maximum eigenvalue $\lambda_{max} = 6.5738$ reported for friendship networks in [5].
Figure 3: The maximum eigenvalue for each one of the 75 villages. The crosses correspond to the actual eigenvalues and the circle to the predicted ones. The blue dotted line corresponds to the maximum eigenvalue of the expectation of the matrix $E A$ of our model. We observe that our over-simplified model manages to predict the bulk of the eigenvalues observed.

However, we do not know the exact network that the authors used, we made a prediction for the maximum eigenvalue of each of the 75 networks in the dataset. The prediction was derived by multiplying the supremum of the support of the spectrum $R(d)$ with the parameter $\sqrt{a(n)} = n^{0.123}$, in order to correct for normalizing. We present the results in Figure 3. We observe that even though each network is relatively small with nodes in the order of hundreds and the model is extremely simple, our asymptotic results show reasonable agreement with the actual values.

References


