The Annals of Probability
2021, Vol. 49, No. 1, 157–179
https://doi.org/10.1214/20-AOP1447

FREENESS OVER THE DIAGONAL FOR LARGE RANDOM MATRICES

BY BENSON AU[1], GUILLAUME CÉBRON[2], ANTOINE DAHLQVIST[3], FRANCK GABRIEL[4] AND CAMILLE MALE[5]

[1]*Department of Mathematics, University of California, San Diego, bau@ucsd.edu*
[2]*IMT; UMR 5219, Université de Toulouse; CNRS, guillaume.cebron@math.univ-toulouse.fr*
[3]*Department of Mathematics, University of Sussex, a.dahlqvist@sussex.ac.uk*
[4]*EPFL-SB-MATH-CSFT, franck.gabriel@epfl.ch*
[5]*Institut de Mathématiques de Bordeaux, Université de Bordeaux, camille.male@math.u-bordeaux.fr*

We prove that independent families of permutation invariant random matrices are asymptotically free with amalgamation over the diagonal, both in expectation and in probability, under a uniform boundedness assumption on the operator norm. We can relax the operator norm assumption to an estimate on sums associated to graphs of matrices, further extending the range of applications (e.g., to Wigner matrices with exploding moments and the sparse regime of the Erdős–Rényi model). The result still holds even if the matrices are multiplied entrywise by random variables satisfying a certain growth condition (e.g., as in the case of matrices with a variance profile and percolation models). Our analysis relies on a modified method of moments based on graph observables.

1. Introduction. Noncommutative (NC) probability is a generalization of classical probability that extends the probabilistic perspective to noncommuting random variables. Following the seminal work of Voiculescu [26], this setting provides a unifying framework for the spectral analysis of random multimatrix models in the large N limit. We outline the basic approach of this program as follows:

1. In the NC framework Voiculescu's *free independence* plays the role of classical independence. This simple parallel yields a surprisingly rich theory with free analogues of many classical concepts (e.g., the free CLT, free cumulants, free entropy and conditional expectations). The scope of free probability further benefits from a robust analytic framework, allowing for a notion of free harmonic analysis [28].
2. Free independence describes the large N limit behavior of classically independent random matrices in many generic situations, notably unitarily invariant ensembles [26]. The free probability machinery then allows for tractable computations of many practical quantities of interest. In particular, one can compute the limiting spectral distribution of rational functions in such matrices [16].

At the same time, many natural random matrix models lie beyond the scope of free probability. One can hope to accommodate such models by defining a suitable new framework. This is the perspective of traffic probability [2, 10, 13–15, 18–20]:

1. The traffic framework adjoins the standard NC probability framework with an operadic structure based on graph observables. The notion of a traffic distribution enriches that of a usual distribution. This additional structure admits a new notion of independence, one that encodes the familiar notions of NC independence.

Received September 2019; revised March 2020.
MSC2020 subject classifications. Primary 15B52, 46L54; secondary 46L53, 60B20.
Key words and phrases. Random matrices, freeness with amalgamation, permutation invariance, traffic probability.

2. Permutation invariant random matrices provide a canonical model of traffic independence in the large N limit.

The assumption of a strong continuous distributional symmetry, such as unitary invariance, ensures that our matrices can be rotated into a so-called "generic position" relative to each other. Free independence then translates this condition into a precise universal rule for calculating the joint asymptotic spectral distribution from the marginals. Naturally, permutation invariance fails to produce the same genericity: for this reason, despite their ubiquity, permutation invariant models traditionally lie outside of the domain of free probability. In what sense then, if any, does permutation invariance determine such a rule?

Surprisingly, contrary to the prevailing intuition, we show that this question can still be answered entirely in terms of the familiar free probability machinery. In this article we show that Voiculescu's notion of a conditional expectation in the context of *operator-valued* free probability provides an analytic framework for traffic independence in the case of large permutation invariant random matrices. Notably, our main result generalizes Voiculescu's celebrated asymptotic freeness theorem for unitarily invariant random matrices to permutation invariant random matrices in the form of freeness with amalgamation over the diagonal.

1.1. *Background.* We begin by recalling the basic framework of operator-valued free probability [27].

DEFINITION 1.1. An *operator-valued probability space* is a triple $(\mathcal{A}, \mathcal{B}, E)$ consisting of a unital algebra $\mathcal{A}$ over $\mathbb{C}$, a unital subalgebra $\mathcal{B} \subset \mathcal{A}$ and a *conditional expectation* $E : \mathcal{A} \to \mathcal{B}$. By a conditional expectation we mean a unital linear $\mathcal{B} - \mathcal{B}$ bimodule map (i.e., $E(b_1 a b_2) = b_1 E(a) b_2$ for any $a \in \mathcal{A}$ and $b_1, b_2 \in \mathcal{B}$).

Let K be an arbitrary index set. We write $\mathcal{B}\langle X_k : k \in K\rangle$ for the free NC algebra generated by $\mathcal{B}$ and the noncommuting indeterminates $(X_k)_{k\in K}$. We call an element of $\mathcal{B}\langle X_k : k \in K\rangle$ a *$\mathcal{B}$-valued polynomial*. In the case of a monomial $M = b_0 X_{k_1} b_1 X_{k_2} \cdots X_{k_n} b_n \in \mathcal{B}\langle X_k : k \in K\rangle$, we define its *degree* $\deg(M) = n$ and its *coefficients* $(b_0, \ldots, b_n)$. The *length* of a $\mathcal{B}$-valued polynomial $P = \sum_{i=1}^{t} M_i$ is the number of monomials t appearing in the finite linear combination defining P. The *operator-valued distribution* (or *E-distribution* for short) of a family $\mathbf{A} = (A^{(k)})_{k\in K} \subset \mathcal{A}$ is the linear map of operator-valued moments

$$E_{\mathbf{A}} : \mathcal{B}\langle X_k : k \in K\rangle \to \mathcal{B}, \qquad P \mapsto E\big[P(\mathbf{A})\big].$$

We say that the families $\mathbf{A}_1 = (A_1^{(k)})_{k\in K}, \ldots, \mathbf{A}_L = (A_L^{(k)})_{k\in K} \subset \mathcal{A}$ are *free with amalgamation over $\mathcal{B}$* (or *free over $\mathcal{B}$* for short) if

$$E\big[\big(P_1(\mathbf{A}_{\ell_1}) - E\big[P_1(\mathbf{A}_{\ell_1})\big]\big) \cdots \big(P_n(\mathbf{A}_{\ell_n}) - E\big[P_n(\mathbf{A}_{\ell_n})\big]\big)\big] = 0$$

for any polynomials $P_1, \ldots, P_n \in \mathcal{B}\langle X_k : k \in K\rangle$ whenever $\ell_1 \neq \ell_2 \neq \cdots \neq \ell_n$.

Note that ordinary freeness is simply the special case of freeness over $\mathcal{B} = \mathbb{C}$. The operator-valued extension retains many of the same properties: for example, the freeness with amalgamation of the families $\mathbf{A}_1, \ldots, \mathbf{A}_L$ uniquely determines the joint operator-valued distribution $E_{\mathbf{A}_1 \sqcup \cdots \sqcup \mathbf{A}_L}$ from the marginal operator-valued distributions $(E_{\mathbf{A}_\ell})_{\ell=1}^{L}$.

EXAMPLE 1.2. Let $\mathcal{M}_N$ denote the algebra of complex $N \times N$ matrices. We define the diagonal map $\Delta : \mathcal{M}_N \to \mathcal{D}_N$ onto the subalgebra $\mathcal{D}_N$ of diagonal matrices by

$$\Delta(A) = \big(\delta_{i,j} A(i,j)\big)_{i,j\in[N]} \quad \forall A = \big(A(i,j)\big)_{i,j\in[N]} \in \mathcal{M}_N.$$

Then, $(\mathcal{M}_N, \mathcal{D}_N, \Delta)$ is an operator-valued probability space.

For random matrices, freeness with amalgamation over the diagonal first appeared in the work of Shlyakhtenko on Gaussian Wigner matrices with a variance profile [24]. Our motivation for considering this question for permutation invariant random matrices comes from a recent work of Boedihardjo and Dykema [8] which proved that random Vandermonde matrices constructed from i.i.d. random variables uniformly distributed on the unit circle are asymptotically $\mathcal{R}$-diagonal over the diagonal matrices. Such matrices are invariant under left multiplication by permutation matrices: if instead this symmetry were with respect to all unitary matrices, then one would obtain convergence to an ordinary scalar-valued $\mathcal{R}$-diagonal element [23], Theorem 15.10. This juxtaposition suggests a link between permutation invariance and freeness with amalgamation over the diagonal.

1.2. *Asymptotic freeness with amalgamation over the diagonal.* In the sequel, when we speak of a family $\mathbf{A}_N$ of random $N \times N$ matrices, we implicitly refer to an entire sequence $(\mathbf{A}_N)_{N\in\mathbb{N}}$ of families $\mathbf{A}_N = (A_N^{(k)})_{k\in K}$ of random $N \times N$ matrices, where K does not depend on N. We say that a family $\mathbf{A}_N$ is *permutation invariant* if for any permutation $\sigma \in \mathfrak{S}_N$, we have the equality in (joint) distribution of the random variables

$$\big(A_N^{(k)}(i,j)\big)_{i,j\in[N],k\in K} \overset{d}{=} \big(A_N^{(k)}(\sigma(i),\sigma(j))\big)_{i,j\in[N],k\in K};$$

or, equivalently, for any $N \times N$ permutation matrix S_N,

$$\big(A_N^{(k)}\big)_{k\in K} \overset{d}{=} \big(S_N A_N^{(k)} S_N^*\big)_{k\in K}.$$

By a *uniform operator norm bound* we mean that $\sup_{(N,k)\in\mathbb{N}\times K} \|A_N^{(k)}\|$ is essentially bounded. We say that a sequence of polynomials $(P_N)_{N\in\mathbb{N}}$ such that $P_N \in \mathcal{D}_N\langle X_k : k \in K\rangle$ satisfies the *uniformly bounded length, degree and coefficient property* (or *LDC* for short) if the lengths of the polynomials are uniformly bounded with uniformly bounded degree for the monomials appearing in the linear combination defining P_N and uniformly bounded operator norms for the coefficients appearing in the monomials. We can now state a simplified version of our main result (Theorem 2.3).

THEOREM 1.3. *Let $\mathbf{A}_{N,1} = (A_{N,1}^{(k)})_{k\in K}, \ldots, \mathbf{A}_{N,L} = (A_{N,L}^{(k)})_{k\in K}$ be independent families of permutation invariant random matrices satisfying a uniform operator norm bound. Then, the families $\mathbf{A}_{N,1}, \ldots, \mathbf{A}_{N,L}$ are asymptotically free with amalgamation over $\mathcal{D}_N$ in the following sense: for any polynomials $P_{N,1}, \ldots, P_{N,n} \in \mathcal{D}_N\langle X_k : k \in K\rangle$ such that the sequences $(P_{N,i})_{N\in\mathbb{N},i\in[n]}$ satisfy the LDC property, the matrix*

$$\varepsilon_N = \Delta\big[\big(P_{N,1}(\mathbf{A}_{N,\ell_1}) - \Delta\big(P_{N,1}(\mathbf{A}_{N,\ell_1})\big)\big)\cdots\big(P_{N,n}(\mathbf{A}_{N,\ell_n}) - \Delta\big(P_{N,n}(\mathbf{A}_{N,\ell_n})\big)\big)\big]$$

converges to zero in (normalized) Schatten p-norm for any $p \in [1,+\infty)$ whenever $\ell_1 \neq \ell_2 \neq \cdots \neq \ell_n$, namely,

$$\lim_{N\to\infty} \mathbb{E}\bigg[\frac{1}{N}\mathrm{Tr}\big[(\varepsilon_N\varepsilon_N^*)^{\frac{p}{2}}\big]\bigg] = 0. \tag{1}$$

Note that the convergence in (1) implies the convergence of the Schatten p-norms $\|\varepsilon_N\|_p \to 0$ in probability as $N \to \infty$. Thus, independent permutation invariant random matrices are asymptotically free over the diagonal in probability in $(\mathcal{M}_N, \mathcal{D}_N, \Delta)$: for a typical realization of $\mathbf{A}_{N,1}, \ldots, \mathbf{A}_{N,L}$, the operator-valued distribution of $\mathbf{A}_{N,1} \sqcup \cdots \sqcup \mathbf{A}_{N,L}$ will be close to the operator-valued distribution of $\widetilde{\mathbf{A}}_{N,1} \sqcup \cdots \sqcup \widetilde{\mathbf{A}}_{N,L}$, where the $(\widetilde{\mathbf{A}}_{N,\ell})_{\ell=1}^L$ are copies of $(\mathbf{A}_{N,\ell})_{\ell=1}^N$ taken to be free with amalgamation over $\mathcal{D}_N$.

In Theorem 2.3 we strengthen this result to the operator-valued probability space generated by $\mathbf{A}_{N,1} \sqcup \cdots \sqcup \mathbf{A}_{N,\ell}$. The remainder of Section 2 is then devoted to generalizations: first,

relaxing the operator norm assumption to an estimate on sums associated to certain graphs (Proposition 2.4); and second, weakening the invariance assumption by allowing the entry-wise multiplication of our matrices by random variables satisfying a certain growth condition (Proposition 2.5).

1.3. *Numerical simulations*. Theorem 2.3 allows us to use the analytic machinery of the operator-valued theory to gain new insight into the spectral analysis of large random matrices. In particular, one can use the operator-valued subordination result of Belinschi, Mai, and Speicher [4] to calculate the limiting spectral distribution of rational functions in our matrices [16].

We illustrate our result by numerically computing the limiting spectral distribution of the sum of two independent Hermitian matrices from various ensembles: GUE matrices with a variance profile; adjacency matrices of sparse Erdős–Rényi graphs; adjacency matrices of percolation on the cycle; and diagonal matrices conjugated by the discrete Fourier transform matrix or unitary Brownian motion.

Organization of the article. We state our main results in Section 2, namely, Theorem 2.3 and its generalizations. In Section 3 we give an algorithm for approximating the operator-valued free convolution over the diagonal and apply it to various matrix models. Finally, Section 4 contains the proofs of the different results stated in Section 2.

2. Statement of results.

2.1. *Freeness with amalgamation for large random matrices*. We work in the natural extension of Example 1.2 to the random setting. In particular, we write $(\mathcal{M}_N(L^{\infty-}), \mathcal{D}_N(L^{\infty-}), \Delta)$ for the operator-valued probability space of random $N \times N$ matrices whose entries have finite moments of all orders. Note that the coefficients of a $\mathcal{D}_N(L^{\infty-})$-valued polynomial are *random* diagonal matrices. In contrast, the papers [8, 24] consider the operator-valued probability space $(\mathcal{M}_N(L^{\infty-}), \mathcal{D}_N, \mathbb{E} \circ \Delta)$. The minimal setting required to formalize Theorem 1.3 in $(\mathcal{M}_N(L^{\infty-}), \mathcal{D}_N(L^{\infty-}), \Delta)$ motivates the following definition.

DEFINITION 2.1. Let $\mathbf{A}_N$ be a family of random matrices in $\mathcal{M}_N(L^{\infty-})$. We define $\mathcal{A}_N \subset \mathcal{M}_N(L^{\infty-})$ to be the smallest unital subalgebra containing $\mathbf{A}_N$ that is closed under the diagonal map Δ. We denote the image of $\mathcal{A}_N$ under the diagonal map by $\mathcal{B}_N = \Delta(\mathcal{A}_N) \subset \mathcal{A}_N$.

Note that $\mathcal{A}_N = \mathcal{B}_N\langle \mathbf{A}_N \rangle$. Furthermore, $\mathcal{B}_N$ is the smallest subalgebra of $\mathcal{D}_N(L^{\infty-})$ such that $\Delta(\mathcal{B}_N\langle \mathbf{A}_N \rangle) \subset \mathcal{B}_N$, and the triple $(\mathcal{A}_N, \mathcal{B}_N, \Delta)$ is an operator-valued probability space. In order to formulate a notion of asymptotic freeness over $\mathcal{B}_N$, the coefficients of the $\mathcal{B}_N$-valued polynomials should somehow be consistent as the dimension N grows. We can encode the coefficients independently of the dimension by using the following notion of a graph monomial from traffic probability [18].

DEFINITION 2.2. A *graph monomial* $g = (G, \eta, \gamma)$ is a finite connected birooted multidigraph $G = (V, E, \mathrm{src}, \mathrm{tar}, v_{\mathrm{in}}, v_{\mathrm{out}})$ together with edge labels $\eta : E \to [L]$ and $\gamma : E \to K$. The maps $\mathrm{src}, \mathrm{tar} : E \to V$ specify the *source* $\mathrm{src}(e)$ and *target* $\mathrm{tar}(e)$ of each edge $e \in E$. We refer to the roots $(v_{\mathrm{in}}, v_{\mathrm{out}}) \in V^2$ as the *input* and the *output*, respectively, though they need not be distinct (in which case we specify that g is a *diagonal graph monomial*). For a graph monomial g we define $\Delta(g)$ as the diagonal graph monomial obtained from g by identifying $v_{\mathrm{in}} \sim v_{\mathrm{out}}$. When convenient, we omit the maps src, tar from the notation and simply abbreviate a multidigraph $G = (V, E)$. See Figure 1 for an illustration.

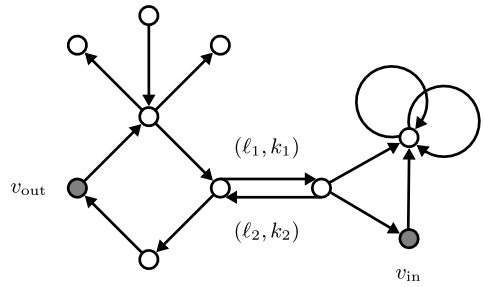


FIG. 1. *A graph monomial with some edge labels indicated.*

For an L-tuple of families $\mathbf{A}_{N,1} = (A_{N,1}^{(k)})_{k\in K}, \ldots, \mathbf{A}_{N,L} = (A_{N,L}^{(k)})_{k\in K}$ of random $N \times N$ matrices, the edge labels η and γ of a graph monomial determine an assignment of each edge $e \in E$ to a matrix $A_{N,\eta(e)}^{(\gamma(e))} \in \mathbf{A}_N = \mathbf{A}_{N,1} \sqcup \cdots \sqcup \mathbf{A}_{N,L}$. This allows us to evaluate a graph monomial g in the family $\mathbf{A}_N$ to obtain a random $N \times N$ matrix $g(\mathbf{A}_N)$ given by the coordinate formula

$$g(\mathbf{A}_N)(i,j) = \sum_{\substack{\phi: V\to[N] \\ \phi(v_{\text{out}})=i, \phi(v_{\text{in}})=j}} \prod_{e\in E} A_{N,\eta(e)}^{(\gamma(e))}\big(\phi(\text{tar}(e)), \phi(\text{src}(e))\big).$$

For convenience, we abbreviate $(\phi(\text{tar}(e)), \phi(\text{src}(e))) = (\phi(e))$. Of course, if $\mathbf{A}_N \subset \mathcal{M}_N(L^{\infty-})$, then $g(\mathbf{A}_N) \in \mathcal{M}_N(L^{\infty-})$.

Note that the action of a diagonal graph monomial produces a diagonal matrix. More generally, we have the identity $\Delta(g(\mathbf{A}_N)) = \Delta(g)(\mathbf{A}_N)$. In particular, if g is the unique graph monomial whose underlying graph consists of a single isolated vertex (with no loops), then $g(\mathbf{A}_N) = I_N$.

In Lemma 4.3 we prove that $\mathcal{B}_N$ is spanned by the matrices $g(\mathbf{A}_N)$ obtained by evaluating $\mathbf{A}_N$ in the so-called *planted cactus-type monomials*, that is, diagonal graph monomials whose underlying graph G is an *oriented cactus* (see Definition 4.1). This implies that a generic $\mathcal{B}_N$-valued monomial takes the form

$$g_{N,0}(\mathbf{A}_N) X_{k_N(1)} g_{N,1}(\mathbf{A}_N) \cdots X_{k_N(d)} g_{N,d_N}(\mathbf{A}_N)$$

for some planted cactus-type monomials $g_{N,j}$ (say with edge set E_j). We define the *full degree* of such a monomial as the sum $d_N + \sum_{j=0}^{d_N} |E_j|$ of the usual degree and the total number of edges appearing in the cacti.

We say that a sequence of polynomials $(P_N)_{N\in\mathbb{N}}$ such that $P_N \in \mathcal{B}_N\langle X_k : k \in K\rangle$ satisfies the *uniformly bounded length and full degree property* (or *LFD* for short) if the lengths of the polynomials are uniformly bounded with uniformly bounded full degree for the monomials appearing in the linear combination defining P_N. This notion is the appropriate generalization of the earlier LDC property to graph monomials and allows us to formulate our main result.

THEOREM 2.3. *Let $\mathbf{A}_{N,1} = (A_{N,1}^{(k)})_{k\in K}, \ldots, \mathbf{A}_{N,L} = (A_{N,L}^{(k)})_{k\in K}$ be independent families of random matrices satisfying a uniform operator norm bound. Assume that each family, except possibly one, is permutation invariant. As before, we write $\mathbf{A}_N = \mathbf{A}_{N,1} \sqcup \cdots \sqcup \mathbf{A}_{N,L}$. Then, $\mathbf{A}_{N,1}, \ldots, \mathbf{A}_{N,L}$ are asymptotically free with amalgamation in $(\mathcal{A}_N, \mathcal{B}_N, \Delta)$ in the following sense: for any polynomials $P_{N,1}, \ldots, P_{N,n} \in \mathcal{B}_N\langle X_k : k \in K\rangle$ such that the sequences $(P_{N,i})_{N\in\mathbb{N}, i\in[n]}$ satisfy the LFD property, the matrix*

$$\varepsilon_N = \Delta\big[\big(P_{N,1}(\mathbf{A}_{N,\ell_1}) - \Delta\big(P_{N,1}(\mathbf{A}_{N,\ell_1})\big)\big) \cdots \big(P_{N,n}(\mathbf{A}_{N,\ell_n}) - \Delta\big(P_{N,n}(\mathbf{A}_{N,\ell_n})\big)\big)\big]$$

converges to zero in Schatten p-norm for any $p \in [1, +\infty)$ *whenever* $\ell_1 \neq \cdots \neq \ell_n$, *namely*,

$$\lim_{N\to\infty} \mathbb{E}\Big[\frac{1}{N}\mathrm{Tr}\big[(\varepsilon_N \varepsilon_N^*)^{\frac{p}{2}}\big]\Big] = 0.$$

Note that we do not assume the convergence of our matrices $\mathbf{A}_N$ in any sense, much as in the context of *deterministic equivalents* [22], Chapter 10.

2.2. *Generalizations*. For starters, we can relax the operator norm assumption to an asymptotic on graph observables.

PROPOSITION 2.4. *The conclusion of Theorem 2.3 still holds if the operator norm bound is replaced by the following weaker assumption*: *for any* $\ell \in [L]$ *and any diagonal graph monomials* $g_1, \ldots, g_n$, *there exists a constant* $C \geq 0$ *such that*

$$\Big|\mathbb{E}\Big[\prod_{i=1}^{n} \mathrm{Tr}\big[g_i(\mathbf{A}_{N,\ell})\big]\Big]\Big| \leq C N^{\sum_{i=1}^n \mathfrak{f}(g_i)/2}, \tag{2}$$

where $\mathfrak{f}(g_i) \geq 2$ *is determined by the forest* F_i *of two-edge connected components of the underlying unlabeled undirected graph* $\underline{G}_i$ *of* g_i (*see Definition 4.13*) *and* C *only depends on* $(\underline{G}_i)_{i=1}^n$.

The assumption in Proposition 2.4 follows from the stronger asymptotic

$$\mathbb{E}\Big[\prod_{i=1}^{n} \mathrm{Tr}\big[g_i(\mathbf{A}_{N,\ell})\big]\Big] = O\big(N^n\big) \tag{3}$$

which is simply the boundedness of the so-called *traffic distribution* [18]. For example, the bound in (3) holds for Wigner matrices with exploding moments and the sparse regime of the Erdős–Rényi model [19], Proposition 4.1, both of which do not satisfy the operator norm assumption [29], Proposition 12. The relevance of the quantity $\mathfrak{f}(g)$ owes to Mingo and Speicher [21], who proved that the bound in (2) holds if the matrices $\mathbf{A}_{N,1}, \ldots, \mathbf{A}_{N,L}$ satisfy a uniform operator norm bound (see Section 4.4).

Our second generalization allows for the entrywise multiplication of our matrices by random variables satisfying a certain growth condition. We will need the notion of a *test graph* $T = (G, \eta, \gamma)$, where $G = (V, E, \mathrm{src}, \mathrm{tar})$ is a finite multidigraph with edge labels $\eta : E \to [L]$ and $\gamma : E \to K$. In contrast to graph monomials, we do not specify any roots for test graphs.

PROPOSITION 2.5. *Let* $\mathbf{A}_{N,1}, \ldots, \mathbf{A}_{N,L}$ *be independent families of permutation invariant random matrices satisfying the asymptotic* (2). *Suppose that*

$$\mathbf{\Gamma}_N = \mathbf{\Gamma}_{N,1} \sqcup \cdots \sqcup \mathbf{\Gamma}_{N,L} = \big(\Gamma_{N,1}^{(k)}\big)_{k\in K} \sqcup \cdots \sqcup \big(\Gamma_{N,L}^{(k)}\big)_{k\in K}$$

is a family of random matrices, independent of $\mathbf{A}_N$, *such that*

$$\Big|\mathbb{E}\Big[\frac{(N-|V|)!}{N!} \sum_{\phi: V \hookrightarrow [N]} \prod_{e\in E} \Gamma_{N,\eta(e)}^{(\gamma(e))}\big(\phi(e)\big)\Big]\Big| \leq C \tag{4}$$

for any test graph $T = (G, \eta, \gamma)$ *with* $G = (V, E)$, *where* $\phi : V \hookrightarrow [N]$ *denotes an injective map and* C *only depends on* $\underline{G}$. *Then, the conclusions of Theorems 1.3 and 2.3 still hold for the families* $\widetilde{\mathbf{A}}_{N,1}, \ldots, \widetilde{\mathbf{A}}_{N,L}$, *where*

$$\widetilde{\mathbf{A}}_{N,\ell} = \big(A_{N,\ell}^{(k)} \circ \Gamma_{N,\ell}^{(k)}\big)_{k\in K}$$

and $\circ$ *denotes the entrywise product*.

Of course, $\frac{(N-|V|)!}{N!}$ is simply the number of maps $\phi : V \hookrightarrow [N]$, so the quantity under consideration in (4) can be written as

$$\left|\mathbb{E}\left[\prod_{e\in E} \Gamma_{N,\eta(e)}^{(\gamma(e))}\big(\Phi(e)\big)\right]\right|$$

for $\Phi : V \hookrightarrow [N]$ a uniformly random injective map independent of $\mathbf{\Gamma}_N$. In particular, this quantity is bounded if the entries of $\mathbf{\Gamma}_N$ are bounded.

We emphasize that the families $\mathbf{\Gamma}_{N,1}, \ldots, \mathbf{\Gamma}_{N,L}$ are not assumed to be independent nor permutation invariant, which subsequently extends to the families $\widetilde{\mathbf{A}}_{N,1}, \ldots, \widetilde{\mathbf{A}}_{N,L}$. For example, Proposition 2.5 applies to Wigner matrices with a variance profile [24].

3. Examples and numerical simulations. In this section we consider pairs of independent random matrices X_N and Y_N sampled from various ensembles covered by our results. For a single realization of our matrices, we compute the empirical spectral distribution of the sum $X_N + Y_N$. For comparison, we use the fixed point algorithm of Belinschi, Mai, and Speicher [4] to compute the spectral density of the free convolution with amalgamation $\Delta_{X_N} \boxplus \Delta_{Y_N}$. Our main result guarantees that the difference between these two computations becomes negligible in the limit, both in expectation and in probability: the simulations provide a visual representation of this convergence in action. The close agreement between the two computations can already be seen from just a single realization of our matrices (as opposed to the average of a large number of simulations) which follows from the convergence in probability form of our result (as opposed to just in expectation).

3.1. *Amalgamated subordination*. To explain the fixed point algorithm, we will need to introduce more of the operator-valued framework.

DEFINITION 3.1. Let $(\mathcal{A}, \mathcal{D}, \Delta)$ be a *C^*-operator-valued probability space*, that is, we further assume that $\mathcal{A}$ is a C^*-algebra, $\mathcal{D}$ is a C^*-subalgebra and Δ is completely positive. The *operator-valued Cauchy transform* of a selfadjoint operator-valued random variable $X \in \mathcal{A}$ is the function

$$G_X : \mathcal{D}^+ \to \mathcal{D}^-, \qquad Z \mapsto \Delta\big[(Z - X)^{-1}\big],$$

where $\mathcal{D}^\pm = \{Z \in \mathcal{D} : \pm\Im(Z) = \pm\frac{Z-Z^*}{2i} > 0\}$. We define the corresponding *H transform*

$$H_X : \mathcal{D}^+ \to \overline{\mathcal{D}^+}, \qquad Z \mapsto G_X(Z)^{-1} - Z.$$

For random matrices the operator-valued Cauchy transform simply corresponds to the diagonal of the matrix resolvent, an important object in random matrix theory. For example, this object has been used to study the adjacency matrices of weighted random graphs [17], CLTs for linear statistics of heavy-tailed random matrices [6] and universality for general Wigner-type matrices [1]. The operator-valued extension further satisfies an analytic subordination property.

THEOREM 3.2 ([4], Theorem 2.7). *Let X and Y be selfadjoint operator-valued random variables in a C^*-operator-valued probability space $(\mathcal{A}, \mathcal{D}, \Delta)$. If X and Y are free over $\mathcal{D}$, then they satisfy the following subordination property*:

$$G_{X+Y}(Z) = G_X\big(\Omega(Z)\big) \quad \forall Z \in \mathcal{D}^+,$$

where $\Omega : \mathcal{D}^+ \to \mathcal{D}^+$ is the unique solution of the fixed point equation

$$\Omega(Z) = F_Z\big(\Omega(Z)\big)$$

for the function

$$F_Z : \mathcal{D}^+ \to \mathcal{D}^+, \qquad \omega \mapsto H_Y\big(H_X(\omega) + Z\big) + Z.$$

Moreover, Ω *can be computed as the iteration*

$$\Omega(Z) = \lim_{n\to\infty} \Omega_n(Z)$$

for $\Omega_{n+1}(Z) = F_Z(\Omega_n(Z))$ *and arbitrary initialization* $\Omega_0(Z) \in \mathcal{D}^+$.

We implement this subordination result to compute an approximation to the operator-valued free convolution. For random matrices X_N and Y_N, the following algorithm generates an approximation to the density $g(x)$ of the operator-valued free convolution $\Delta_{X_N} \boxplus \Delta_{Y_N}$ at the values $x \in \mathbb{R}$ where such a density exists:

1. Simulate a realization of X_N and Y_N.
2. Set $\Omega_0(Z) = (x + i\varepsilon)I_N \in \mathcal{D}_N$ for a small value of $\varepsilon > 0$. Compute the terms of the sequence $(\Omega_n(Z))_{n\in\mathbb{N}}$,

$$\Omega_{n+1}(Z) = F_Z\big(\Omega_n(Z)\big) = H_{Y_N}\big(H_{X_N}\big(\Omega_n(Z)\big) + Z\big) + Z,$$

until the norm $\|\Omega_{n+1}(Z) - \Omega_n(Z)\|$ is less than a prescribed threshold for some value of $n = n_0$.
3. The value of the density $g(x)$ can then be approximated by

$$-\frac{1}{\pi}\Im\left(\frac{1}{N}\mathrm{Tr}\big[\big(\Omega_{n_0}(Z) - X_N\big)^{-1}\big]\right),$$

provided ε is sufficiently small and the distribution admits a density at x.

REMARK 3.3. We mention two possible extensions of this algorithm:

1. The operator-valued R-transform of Y is the map $R_Y : \mathcal{D}^+ \to \mathcal{D}^+$ uniquely determined by the relation $G_Y(Z) = (Z - R_Y(G_Y(Z)))^{-1}$. In fact, the function $\Omega(Z)$ in Theorem 3.2 equals $Z - R_Y(G_{X+Y}(Z))$. Knowledge of the operator-valued R-transform of either X_N or Y_N provides faster algorithms that do not require the simulation of the matrices X_N and Y_N [22], Theorem 11 of Chapter 9.
2. The fixed point algorithm described by Belinschi, Mai, and Speicher [4] can also be used to compute the distribution of NC *rational functions* in X and Y through an appropriate linearization [16].

3.2. *Matrix models.* We now define the various matrix models in our simulations:

1. We write $\mathrm{GUE}_{\mathrm{vp}}(N, \eta)$ for the random matrix with block variance profile

$$\sqrt{\frac{8}{5\eta + 3}}\begin{pmatrix} \sqrt{\eta}X_{1,1} & X_{1,2} \\ X_{2,1} & \sqrt{\eta}X_{2,2} \end{pmatrix},$$

where $\eta > 0$ is a parameter; $X_{1,1}$ and $X_{2,2}$ are square matrices of order $N/4$ and $3N/4$, respectively; and

$$\begin{pmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \end{pmatrix}$$

is a normalized GUE matrix of order $N \in 4\mathbb{N}$.

2. We write $\mathrm{ER}(N,d)$ for the standardized adjacency matrix of a sparse Erdős–Rényi graph, namely,

$$\mathrm{ER}(N,d)=\frac{Y_N(d)-\frac{d}{N-1}(\mathbb{J}_N-I_N)}{\sqrt{d(1-d(N-1)^{-1})}},$$

where $Y_N(d)$ is a random real symmetric $N\times N$ matrix with zeros on the diagonal and i.i.d. Bernoulli random variables with parameter $\frac{d}{N-1}$ otherwise (up to the symmetry constraint); $\mathbb{J}_N$ is the all-ones matrix of order N; I_N is the identity matrix of order N, and $d>0$ is a parameter.

3. We write $\mathrm{Perm}(N)$ for the random real symmetric $N\times N$ matrix

$$\frac{1}{\sqrt{2}}(V_N+V_N^*-2\mathbb{J}_N),$$

where V_N is a uniformly distributed permutation matrix of order N. We further write $\Gamma(N,p)$ for the random real symmetric $N\times N$ matrix with zeros on the diagonal and i.i.d. Bernoulli random variables with parameter p otherwise (up to the symmetry constraint), where $\Gamma(N,p)$ and $\mathrm{Perm}(N)$ are independent. This allows us to define the percolated model

$$\mathrm{Perm}(N,p)=\frac{1}{\sqrt{p}}\Gamma(N,p)\circ\mathrm{Perm}(N).$$

4. We define the $N\times N$ diagonal matrices

$$D_{N,1}=\mathrm{diag}(1,-1,1,-1,\dots,1,-1);$$

$$D_{N,2}=\mathrm{diag}(\underbrace{-1,\dots,-1}_{N/2},\underbrace{1,\dots,1}_{N/2});$$

$$D_{N,3}=\sqrt{\frac{3}{14}}\,\mathrm{diag}\Big(\underbrace{-2+\frac{2}{N},-2+\frac{4}{N},\dots,-1}_{N/2},\underbrace{1+\frac{2}{N},1+\frac{4}{N},\dots,2}_{N/2}\Big),$$

where we assume that $N\in 2\mathbb{N}$.

5. For a given $N\times N$ diagonal matrix D_N, we write $\mathrm{FFT}_{D_N}(N,p)$ for the random $N\times N$ matrix

$$\frac{1}{\sqrt{p}}\Gamma(N,p)\circ\big(V_NU_ND_NU_N^*V_N^*\big),$$

where $\Gamma(N,p)$ and V_N are as before and U_N is the $N\times N$ discrete Fourier transform matrix

$$U_N(j,k)=\frac{1}{\sqrt{N}}e^{-2\pi i(j-1)(k-1)}.$$

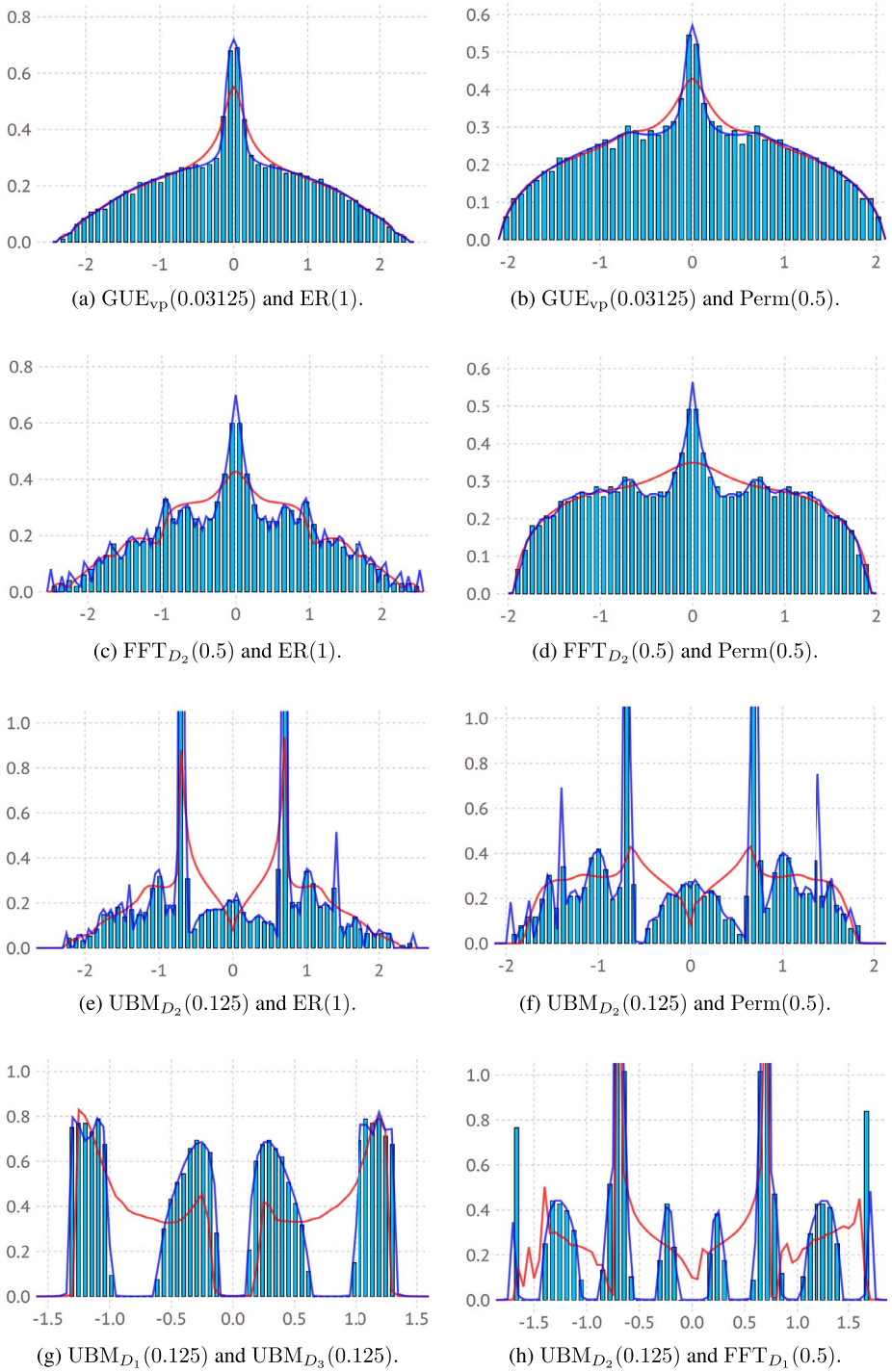
We refer the reader to [12] for earlier work and the motivation behind this model.

6. For a given $N\times N$ diagonal matrix D_N, we write $\mathrm{UBM}_{D_N}(N,p)$ for the random $N\times N$ matrix

$$U_{N,t}D_NU_{N,t}^*,$$

where $U_{N,t}$ is a unitary Brownian motion on the unitary group of order N at time t. We refer the reader to [7] for the definition of unitary Brownian motion and the associated notion of t-freeness.

REMARK 3.4. Heavy-tailed models also fit within our framework under a suitable truncation. We refer the reader to the works [5, 9, 11] for further reading and particularly [3] for an instance of such an operator-valued convolution.



(a) $\mathrm{GUE}_{\mathrm{vp}}(0.03125)$ and $\mathrm{ER}(1)$.

(b) $\mathrm{GUE}_{\mathrm{vp}}(0.03125)$ and $\mathrm{Perm}(0.5)$.

(c) $\mathrm{FFT}_{D_2}(0.5)$ and $\mathrm{ER}(1)$.

(d) $\mathrm{FFT}_{D_2}(0.5)$ and $\mathrm{Perm}(0.5)$.

(e) $\mathrm{UBM}_{D_2}(0.125)$ and $\mathrm{ER}(1)$.

(f) $\mathrm{UBM}_{D_2}(0.125)$ and $\mathrm{Perm}(0.5)$.

(g) $\mathrm{UBM}_{D_1}(0.125)$ and $\mathrm{UBM}_{D_3}(0.125)$.

(h) $\mathrm{UBM}_{D_2}(0.125)$ and $\mathrm{FFT}_{D_1}(0.5)$.

FIG. 2. *Results of the numerical simulations.*

3.3. *Simulation parameters.* Figure 2 reports the results of our simulations for the various models of X_N and Y_N indicated in the captions. In each case, we record:

- in light blue, the histogram of eigenvalues for one realization of the sum $\frac{1}{\sqrt{2}}(X_N + Y_N)$ for $N = 1000$ (we omit the parameter N in the captions);
- in blue, the density of the operator-valued free convolution $\Delta_{X_N} \boxplus \Delta_{Y_N}$ over the diagonal;

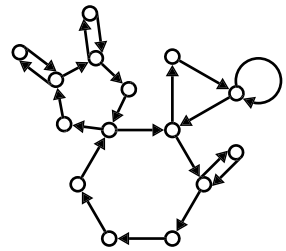


FIG. 3. *An oriented cactus with seven simple directed cycles.*

- in red, the density of the free convolution $\mu_{X_N} \boxplus \mu_{Y_N}$.

In the algorithm we choose the small parameter $\varepsilon = 0.001$. We also stop the fixed point iteration at the threshold $\|\Omega_{n+1}(Z) - \Omega_n(Z)\| \leq 0.001$.

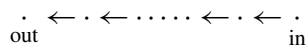
The prediction by freeness over the diagonal accurately fits the histogram in each of the situations we considered. In cases (a)–(d) we see that the deviation between freeness and freeness with amalgamation is mainly at the center of the spectrum. In cases (e)–(h) the time of the unitary Brownian motion is quite small, and so the difference between freeness and freeness with amalgamation is more substantial.

4. Proofs of the main results.

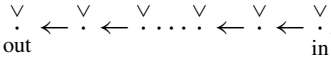
4.1. *Graph polynomial algebras.* We first define the notion of a cactus graph.

DEFINITION 4.1. We say that a finite connected multidigraph G is a *cactus* if every edge belongs to a unique simple cycle. In the case that each such cycle is directed, then we further specify that G is an *oriented cactus*. See Figure 3 for an illustration.

We write $\mathcal{C}$ for the set of all *planted cactus-type monomials*, that is, diagonal graph monomials whose underlying graph is an oriented cactus. We consider the action of these graph monomials on our matrices $\mathbf{A}_N$: in particular, we write $\mathcal{C}_N$ for the vector space generated by $(g(\mathbf{A}_N))_{g\in\mathcal{C}}$. More generally, we define $\mathcal{F}$ to be the set of all graph monomials obtained by starting with a directed path



and attaching oriented cacti to the vertices of this path



Similarly, we write $\mathcal{F}_N$ for the vector space generated by $(g(\mathbf{A}_N))_{g\in\mathcal{F}}$. Note that $\mathcal{C} \subset \mathcal{F}$, and so $\mathcal{C}_N \subset \mathcal{F}_N$. Furthermore, $\mathcal{F}_N = \mathcal{C}_N\langle \mathbf{A}_N\rangle$, namely, $\mathcal{F}_N$ is the vector space generated by elements of the form

$$P = g_0(\mathbf{A}_N)A_{N,\ell_1}^{(k_1)} g_1(\mathbf{A}_N)\cdots A_{N,\ell_d}^{(k_d)} g_d(\mathbf{A}_N), \quad g_i \in \mathcal{C}. \tag{5}$$

Indeed, this follows from the fact that such an element P can be written as $g(\mathbf{A}_N)$ for the graph monomial $g \in \mathcal{F}$ given by

$$g = \overset{g_0}{\underset{\text{out}}{\cdot}} \xleftarrow{(\ell_1,k_1)} \overset{g_1}{\cdot} \xleftarrow{(\ell_2,k_2)} \overset{g_2}{\cdot} \cdots \overset{g_{d-1}}{\cdot} \xleftarrow{(\ell_d,k_d)} \overset{g_d}{\underset{\text{in}}{\cdot}}. \tag{6}$$

LEMMA 4.2. *The vector spaces $\mathcal{C}_N$ and $\mathcal{F}_N$ are unital algebras with $\Delta(\mathcal{F}_N) = \mathcal{C}_N$.*

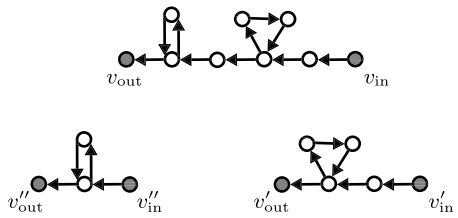


FIG. 4. *An example of factoring when* $v_{\text{out}} \neq v_{\text{in}}$.

PROOF. For any graph monomials g_1 and g_2, the product $g_1(\mathbf{A}_N)g_2(\mathbf{A}_N)$ of matrices equals $g(\mathbf{A}_N)$ for the graph monomial g obtained from g_1 and g_2 by concatenation. In particular, one obtains g from the disjoint union of g_1 and g_2 by identifying the input of g_1 with the output of g_2 and forgetting their distinguished roles. It follows that both $\mathcal{C}_N$ and $\mathcal{F}_N$ are algebras. Moreover, recall that the action of the trivial graph monomial $g = \cdot_{\text{in/out}} \in \mathcal{C}$ produces the identity matrix $g(\mathbf{A}_N) = I_N$.

For the second statement, recall that $\Delta(g(\mathbf{A}_N)) = \Delta(g)(\mathbf{A}_N)$. The result then follows from the simple fact that $\Delta(\mathcal{F}) = \mathcal{C}$. □

Lemma 4.2 allows us to characterize the algebras appearing in Definition 2.1.

LEMMA 4.3. *We have the equality* $\mathcal{F}_N = \mathcal{A}_N$ *(and hence* $\mathcal{C}_N = \mathcal{B}_N$*).*

PROOF. Recall that $\mathcal{A}_N$ is the smallest unital subalgebra containing $\mathbf{A}_N$ that is closed under the diagonal map Δ, and $\mathcal{B}_N = \Delta(\mathcal{A}_N)$. By Lemma 4.2 it suffices to prove that $\mathcal{F}_N = \mathcal{A}_N$. Note that one direction is immediate: by construction, $\mathbf{A}_N \subset \mathcal{F}_N$ and $\Delta(\mathcal{F}_N) \subset \mathcal{F}_N$, whence $\mathcal{A}_N \subset \mathcal{F}_N$.

We prove the reverse inclusion by induction on the number of edges of a graph monomial $g \in \mathcal{F}$. If g has a single edge, then $g(\mathbf{A}_N) \in \mathbf{A}_N \cup \Delta(\mathbf{A}_N) \subset \mathcal{A}_N$. Now, assume that for some $n \geq 2$, every graph monomial $g \in \mathcal{F}$ with fewer than n edges produces a matrix $g(\mathbf{A}_N) \in \mathcal{A}_N$. Consider an $h \in \mathcal{F}$ with exactly n edges: if the input and the output of h are not equal or if they are equal but belong to more than one cycle, then we can factor $h(\mathbf{A}_N) = h_1(\mathbf{A}_N)h_2(\mathbf{A}_N)$ for some graph monomials $h_1, h_2 \in \mathcal{F}$ with fewer than n edges (Figure 4). Since $\mathcal{A}_N$ is an algebra, it follows that $h(\mathbf{A}_N) \in \mathcal{A}_N$.

If instead h is a diagonal graph monomial such that the common root v belongs to a unique cycle, then we can split the vertex v into two vertices $v'_{\text{out}} \neq v'_{\text{in}}$ (Figure 5). This allows us to construct a product P of the form (5) such that $\Delta(P) = h(\mathbf{A}_N)$. We can then factor $P = h_1(\mathbf{A}_N)h_2(\mathbf{A}_N)$ as before, where $h_1(\mathbf{A}_N), h_2(\mathbf{A}_N) \in \mathcal{A}_N$ by the induction hypothesis. We conclude that $h(\mathbf{A}_N) = \Delta(h_1(\mathbf{A}_N)h_2(\mathbf{A}_N)) \in \mathcal{A}_N$. □

4.2. *Some preliminary lemmas.* Recall that a sequence of polynomials $(P_N)_{N\in\mathbb{N}}$ such that $P_N \in \mathcal{B}_N\langle X_k : k \in K\rangle$ satisfies the LFD property if the lengths of the polynomials P_N are uniformly bounded with uniformly full degree for the monomials appearing in the linear

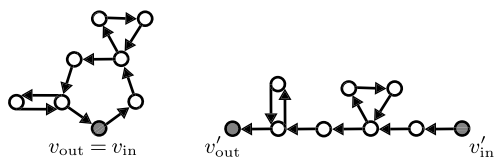


FIG. 5. *An example of splitting when* $v_{\text{out}} = v_{\text{in}}$ *belongs to a unique cycle.*

combination defining P_N. Thus, to prove Theorem 2.3 and its generalizations, it suffices to consider the case of monomials $P_{N,1}, \dots, P_{N,n}$ of uniformly bounded full degree. Indeed, our uniform bound on the lengths allows us to use multilinearity to extend the result to the general case. Furthermore, the uniform boundedness of the full degree for our $\mathcal{B}_N$-valued monomials

$$P_{N,i} = g_{N,i,0}(\mathbf{A}_N) X_{k_{N,i}(1)} g_{N,i,1}(\mathbf{A}_N) \cdots X_{k_{N,i}(d_{N,i})} g_{N,i,d_{N,i}}(\mathbf{A}_N)$$

ensures that the number of underlying unlabeled undirected graphs $\underline{G}_{N,i,j}$ appearing in $(g_{N,i,j})_{N\in\mathbb{N}, i\in[n], j\in[d_{N,i}]}$ is finite. The uniformity of our asymptotic (2) then allows us to further restrict to the case of fixed monomials $P_{N,i}$ in the sense that $d_{N,i} \equiv d_i$ and $\underline{G}_{N,i,j}$ do not depend on N.

We start by considering arbitrary families of random matrices $\mathbf{A}_{N,1}, \dots, \mathbf{A}_{N,L}$. We can assume that the families are each closed under the adjoint by enlarging them if necessary. Our first lemma reduces the even case $p = 2q \in 2\mathbb{N}$ of our main theorem to a single limit.

LEMMA 4.4. *In the notation of Theorem 2.3, if*

$$\lim_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\operatorname{Tr}(\varepsilon_N)\right] = 0$$

whenever

$$\varepsilon_N = \Delta\big[\big(P_{N,1}(\mathbf{A}_{N,\ell_1}) - \Delta\big(P_{N,1}(\mathbf{A}_{N,\ell_1})\big)\big) \cdots \big(P_{N,n}(\mathbf{A}_{N,\ell_n}) - \Delta\big(P_{N,n}(\mathbf{A}_{N,\ell_n})\big)\big)\big]$$

for some $\ell_1 \neq \ell_2 \neq \cdots \neq \ell_n$ and

$$P_{N,i} = g_{N,i,0}(\mathbf{A}_N) X_{k_{N,i}(1)} g_{N,i,1}(\mathbf{A}_N) \cdots X_{k_{N,i}(d_i)} g_{N,i,d_i}(\mathbf{A}_N), \quad g_{N,i,j} \in \mathcal{C},$$

with uniformly bounded full degree, then

$$\lim_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\operatorname{Tr}\big[(\varepsilon_N \varepsilon_N^*)^q\big]\right] = 0$$

for any such ε_N and $q \in \mathbb{N}$ as well.

PROOF. Note that $(\varepsilon_N\varepsilon_N^*)^q = \varepsilon_N\varepsilon_N^*(\varepsilon_N\varepsilon_N^*)^{q-1}$ can be written as

$$\Delta\big[\big(P_{N,1}(\mathbf{A}_{N,\ell_1}) - \Delta\big(P_{N,1}(\mathbf{A}_{N,\ell_1})\big)\big) \cdots \big(P_{N,n}(\mathbf{A}_{N,\ell_n}) - \Delta\big(P_{N,n}(\mathbf{A}_{N,\ell_n})\big)\big)\big]\varepsilon_N^*\big(\varepsilon_N\varepsilon_N^*\big)^{q-1},$$

where $\varepsilon_N^*(\varepsilon_N\varepsilon_N^*)^{q-1} \in \mathcal{B}_N$. Lemma 4.3 then implies that

$$\varepsilon_N^*\big(\varepsilon_N\varepsilon_N^*\big)^{q-1} = \sum_{l=1}^{t_N} c_{N,l} h_{N,\ell}(\mathbf{A}_N)$$

for some coefficients $c_{N,l} \in \mathbb{C}$ and graph monomials $h_{N,l} \in \mathcal{C}$ (say with edge set $E_{N,l}$). At the same time, the bound on the full degree guarantees that

$$\sup_{N\in\mathbb{N}} t_N < \infty \quad \text{and} \quad \sup_{N\in\mathbb{N}} \sum_{l=1}^{t_N} |E_{N,l}| < \infty.$$

So, it suffices to prove that

$$\lim_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\operatorname{Tr}\big[\varepsilon_N h_N(\mathbf{A}_N)\big]\right] = 0$$

for any sequence $h_N \in \mathcal{C}$ with edge set E_N such that $\sup_{N\in\mathbb{N}} |E_N| < \infty$.

Now, since Δ is a conditional expectation, we know that

$$\Delta\big(P_{N,n}(\mathbf{A}_{N,\ell_n})\big)h_N(\mathbf{A}_N) = \Delta\big(P_{N,n}(\mathbf{A}_{N,\ell_n})h_N(\mathbf{A}_N)\big).$$

We can then modify the last coefficient of the monomial $P_{N,n}$ to include the factor of $h_N(\mathbf{A}_N)$. In particular, we can replace

$$P_{N,n} = g_{N,n,0}(\mathbf{A}_N)X_{k_{N,n}(1)}g_{N,n,1}(\mathbf{A}_N)\cdots X_{k_{N,n}(d_n)}g_{N,n,d_n}(\mathbf{A}_N)$$

with the monomial

$$\widetilde{P}_{N,n} = g_{N,n,0}(\mathbf{A}_N)X_{k_{N,n}(1)}g_{N,n,1}(\mathbf{A}_N)\cdots X_{k_{N,n}(d_n)}\widetilde{g}_{N,n,d_n}(\mathbf{A}_N),$$

where $\widetilde{g}_{N,n,d_n}(\mathbf{A}_N) = g_{N,n,d_n}(\mathbf{A}_N)h_N(\mathbf{A}_N)$. This new sequence $(\widetilde{P}_{N,n})_{N\in\mathbb{N}}$ still has uniformly bounded full degree since the number of edges we are adding via h_N is uniformly bounded. In particular, we can apply the assumption in the statement of the lemma to

$$\widetilde{\varepsilon}_N = \Delta\big[\big(P_{N,1}(\mathbf{A}_{N,\ell_1}) - \Delta\big(P_{N,1}(\mathbf{A}_{N,\ell_1})\big)\big)\cdots\big(\widetilde{P}_{N,n}(\mathbf{A}_{N,\ell_n}) - \Delta\big(\widetilde{P}_{N,n}(\mathbf{A}_{N,\ell_n})\big)\big)\big]$$

which shows that $\lim_{N\to\infty}\mathbb{E}[\frac{1}{N}\mathrm{Tr}(\widetilde{\varepsilon}_N)] = 0$. The result then follows from the observation that $\widetilde{\varepsilon}_N = \varepsilon_N h_N(\mathbf{A}_N)$. □

Furthermore, the even case of our main theorem implies the general case.

LEMMA 4.5. *In the notation of Theorem 2.3, if*

$$\lim_{N\to\infty}\mathbb{E}\bigg[\frac{1}{N}\mathrm{Tr}\big[(\varepsilon_N\varepsilon_N^*)^{\frac{p}{2}}\big]\bigg] = 0$$

for any $p\in 2\mathbb{N}$, then the result also holds for any $p\in[1,+\infty)$.

PROOF. By construction, ε_N is a diagonal random matrix. In particular, Theorem 2.3 for $p = 2q\in 2\mathbb{N}$ can be rephrased as

$$\mathbb{E}\big[\mathrm{Tr}\big[(\varepsilon_N\varepsilon_N^*)^q\big]\big] = \mathbb{E}\bigg[\sum_{i=1}^N|\varepsilon_N(i,i)|^p\bigg] = \|\varepsilon_N\|_p^p = o(N), \tag{7}$$

where $\|\cdot\|_p$ denotes the p-norm of a random vector in $\mathbb{C}^N$.

Let us consider $r\in[1,+\infty)$. For any $p\in 2\mathbb{N}$ such that $p\geq r$, Hölder's inequality and equation (7) imply that

$$\begin{aligned}\|\varepsilon_N\|_r^r &= \mathbb{E}\bigg[\sum_{i=1}^N|\varepsilon_N(i,i)|^r\bigg] \leq \bigg(\mathbb{E}\bigg[\sum_{i=1}^N\big(|\varepsilon_N(i,i)|^r\big)^{\frac{p}{r}}\bigg]\bigg)^{\frac{r}{p}}\bigg(\sum_{i=1}^N|1|^{\frac{p}{p-r}}\bigg)^{\frac{p-r}{p}}\\ &= \big(\|\varepsilon_N\|_p^p\big)^{\frac{r}{p}}N^{1-\frac{r}{p}} = o\big(N^{\frac{r}{p}}\big)N^{1-\frac{r}{p}} = o(N).\end{aligned}$$

□

For any $i\in[n]$, let $P_{N,i}\in\mathcal{B}_N\langle X_k : k\in K\rangle$ be a $\mathcal{B}_N$-valued monomial. We write $M_{N,i} = P_{N,i}(\mathbf{A}_{N,\ell_i})$ and $\mathring{M}_{N,i} = M_{N,i} - \Delta(M_{N,i})$, where we recall that $\ell_1\neq\ell_2\neq\cdots\neq\ell_n$. To compute the trace $\mathbb{E}[\frac{1}{N}\mathrm{Tr}(\mathring{M}_{N,1}\cdots\mathring{M}_{N,n})]$, we will need a method of moments adapted to graph observables, which we introduce now.

DEFINITION 4.6. Recall that a *test graph* $T = (G, \eta, \gamma)$ is a finite multidigraph $G = (V, E, \mathrm{src}, \mathrm{tar})$ with edge labels $\eta : E \to [L]$ and $\gamma : E \to K$. For a test graph T we define the quantities:

$$\tau_N[T] = \mathbb{E}\Big[\frac{1}{N^{c(T)}} \sum_{\phi: V \to [N]} \prod_{e \in E} A_{N,\eta(e)}^{(\gamma(e))}(\phi(e))\Big]; \tag{8}$$

$$\tau_N^0[T] = \mathbb{E}\Big[\frac{1}{N^{c(T)}} \sum_{\phi: V \hookrightarrow [N]} \prod_{e \in E} A_{N,\eta(e)}^{(\gamma(e))}(\phi(e))\Big], \tag{9}$$

where $c(T)$ counts the number of connected components of T, $\phi : V \hookrightarrow [N]$ denotes an injective map and $(\phi(e)) = (\phi(\mathrm{tar}(e)), \phi(\mathrm{src}(e)))$.

For any partition $\pi \in \mathcal{P}(V)$, let $\beta_\pi : V \to \pi$ be the map that sends each vertex in V to its block in π. We define $T^\pi = (G^\pi, \eta^\pi, \gamma^\pi)$ as the test graph obtained from T by identifying the vertices in each block of π. Formally, $V^\pi = \pi$; $E^\pi = E$; $\mathrm{src}^\pi = \beta_\pi \circ \mathrm{src}$; $\mathrm{tar}^\pi = \beta_\pi \circ \mathrm{tar}$; $\eta^\pi = \eta$; and $\gamma^\pi = \gamma$. We say that T^π is a *quotient* of T. One can then relate the quantities (8) and (9) by [18], Lemma 2.6:

$$\tau_N[T] = \sum_{\pi \in \mathcal{P}(V)} N^{c(T^\pi) - c(T)} \tau_N^0[T^\pi]; \tag{10}$$

$$\tau_N^0[T] = \sum_{\pi \in \mathcal{P}(V)} N^{c(T^\pi) - c(T)} \tau_N[T^\pi] \mathrm{M\,\ddot{o}b}(0_V, \pi), \tag{11}$$

where 0_V is the singleton partition and $\mathrm{M\ddot{o}b}(0_V, \pi) = \prod_{B \in \pi} (-1)^{|B|-1}(|B| - 1)!$ is the Möbius function on the poset of partitions [25], Example 3.10.4.

REMARK 4.7. We allow our test graphs to be disconnected in this article, in contrast to the convention in [18]. The general case follows much as in the usual situation.

The expansion

$$M_{N,i} = P_{N,i}(\mathbf{A}_{N,\ell_i}) = g_{N,i,0}(\mathbf{A}_N) A_{N,\ell_i}^{(k_{N,i}(1))} g_{N,i,1}(\mathbf{A}_N) \cdots A_{N,\ell_i}^{(k_{N,i}(d_i))} g_{N,i,d_i}(\mathbf{A}_N)$$

determines a graph monomial $t_{N,i} \in \mathcal{F}$ such that $t_{N,i}(\mathbf{A}_N) = M_{N,i}$, namely,

$$t_{N,i} = \overset{g_{N,i,0}}{\underset{\text{out}}{\bullet}} \xleftarrow{(\ell_i, k_{N,i}(1))} \overset{g_{N,i,1}}{\bullet} \xleftarrow{(\ell_i, k_{N,i}(2))} \overset{g_{N,i,2}}{\bullet} \cdots \overset{g_{N,i,d_i-1}}{\bullet} \xleftarrow{(\ell_i, k_{N,i}(d_i))} \overset{g_{N,i,d_i}}{\underset{\text{in}}{\bullet}}. \tag{12}$$

Formally, t_i consists of a directed path of length d_i starting from the right at the input and terminating at the left at the output with: edges labeled in decreasing order $(\ell_i, k_{N,i}(d_i)), \ldots, (\ell_i, k_{N,i}(0))$ along the direction of this path; and planted cactus-type monomials attached to the vertices of this path by gluing the common root of g_{N,i,d_i-j+1} to the jth vertex along this path.

Let $T_N = (G_N, \eta_N, \gamma_N)$ be the test graph obtained by identifying the output of $t_{N,i}$ with the input of $t_{N,i-1}$ for $i \in [n]$ with the convention that $t_{N,0} = t_{N,n}$. In particular, T_N inherits the edge labels from each $t_{N,i}$. Note that $T_N = \Delta(t_{N,1} \cdots t_{N,n})$ as unrooted graphs. One can easily verify that $\mathbb{E}[\frac{1}{N}\mathrm{Tr}(M_{N,1} \cdots M_{N,n})] = \tau_N[T_N]$.

For each $i \in [n]$, we can view $t_{N,i}$ as a subgraph of T_N. Note that the underlying graph $\underline{G}_N = (V, E)$ of G_N does not depend on N since d_i is constant and the underlying graphs $\underline{G}_{N,i,j}$ of the $g_{N,i,j}$ do not depend on N. So, we can write $v_{N,i} \equiv v_i$ for the vertex in T_N corresponding to the input of $t_{N,i}$ with the convention that $v_0 = v_n$. The following lemma describes the cancellations in equation (10) when one replaces $M_{N,i}$ with $\mathring{M}_{N,i} = M_{N,i} - \Delta(M_{N,i})$.

LEMMA 4.8. *With the notation above for the test graph T_N, we have the equality*

$$\mathbb{E}\left[\frac{1}{N}\mathrm{Tr}(\mathring{M}_{N,1}\cdots\mathring{M}_{N,n})\right] = \sum_{\substack{\pi\in\mathcal{P}(V)\text{ s.t.}\\ v_i\nsim_\pi v_{i-1}\ \forall i\in[n]}} \tau_N^0[T_N^\pi]. \tag{13}$$

PROOF. For any $I\subset[n]$, let $T_{N,I}$ denote the test graph obtained from T_N by identifying the input of $t_{N,i}$ with the output of $t_{N,i}$ for each $i\in I$. Then, by definition of the $\mathring{M}_{N,i}$,

$$\mathbb{E}\left[\frac{1}{N}\mathrm{Tr}(\mathring{M}_{N,1}\cdots\mathring{M}_{N,n})\right] = \sum_{I\subset[n]}(-1)^{|I|}\tau_N[T_{N,I}].$$

Using equation (10), we can rewrite this as

$$\mathbb{E}\left[\frac{1}{N}\mathrm{Tr}(\mathring{M}_{N,1}\cdots\mathring{M}_{N,n})\right] = \sum_{I\subset[n]}(-1)^{|I|}\sum_{\substack{\pi\in\mathcal{P}(V)\text{ s.t.}\\ v_i\sim_\pi v_{i-1}\ \forall i\in I}}\tau_N^0[T_N^\pi].$$

For any $\pi\in\mathcal{P}(V)$, let $J_\pi=\{i\in[n]: v_i\sim_\pi v_{i-1}\}$. Interchanging the order of the summation, we obtain

$$\mathbb{E}\left[\frac{1}{N}\mathrm{Tr}(\mathring{M}_{N,1}\cdots\mathring{M}_{N,n})\right] = \sum_{\pi\in\mathcal{P}(V)}\Bigg(\sum_{I\subset J_\pi}(-1)^{|I|}\Bigg)\tau_N^0[T_N^\pi].$$

The result then follows from the fact that the sum in the parentheses vanishes whenever $J_\pi\neq\varnothing$. □

Our analysis of each of the remaining terms $\tau_N^0[T_N^\pi]$ in Lemma 4.8 relies on the geometry of a graph that we introduce now.

DEFINITION 4.9. A *colored component* of a test graph $T=(V,E,\eta,\gamma)$ is a connected component in the subtest graph $T_{\{\ell\}\times K}$ obtained from T by only keeping the edges e with label $(\eta(e),\gamma(e))\in\{\ell\}\times K$, where $\ell\in[L]$. In particular, evaluating $\tau_N[C]$ for a colored component C, say of color $\{\ell\}\times K$, only involves the matrices in the family $\mathbf{A}_{N,\ell}\subset\mathbf{A}_N=\mathbf{A}_{N,1}\sqcup\cdots\sqcup\mathbf{A}_{N,L}$. We write $\mathcal{CC}(T)$ for the set of colored components of T.

We define the *graph of colored components* $\mathcal{GCC}(T)=(\mathcal{V},\mathcal{E})$ as the simple bipartite graph with vertex set $\mathcal{V}=\mathcal{CC}(T)\sqcup V$ and edges determined by inclusion: $v\sim_{\mathcal{E}} C$ if $v\in V$ is a vertex of the colored component $C\in\mathcal{CC}(T)$.

REMARK 4.10. Our definition of the graph of colored components slightly differs from the original one in [18], Definition 2.10, where one discards the trivial colored components in $\mathcal{CC}(T)$ consisting of an isolated vertex and only considers the vertices in V that belong to more than one of the remaining colored components.

We return to the test graph $T_N=(G_N,\eta_N,\gamma_N)$ considered in Lemma 4.8.

LEMMA 4.11. *None of the partitions π in (13) result in a tree for the graph of colored components $\mathcal{GCC}(T_N^\pi)=(\mathcal{V}_{N,\pi},\mathcal{E}_{N,\pi})$.*

PROOF. Suppose, for a contradiction, that there exists a partition $\pi\in\mathcal{P}(V)$ such that $v_i\nsim_\pi v_{i-1}$ for all $i\in[n]$ with $\mathcal{GCC}(T_N^\pi)$ a tree. Let c_N be the simple directed cycle subgraph in T_N defined by traversing the edges of the directed paths at the base of each $t_{N,i}$; see (12). Similarly, we define $C_{N,i}\in\mathcal{CC}(T_N^\pi)$ as the colored component containing the edges of

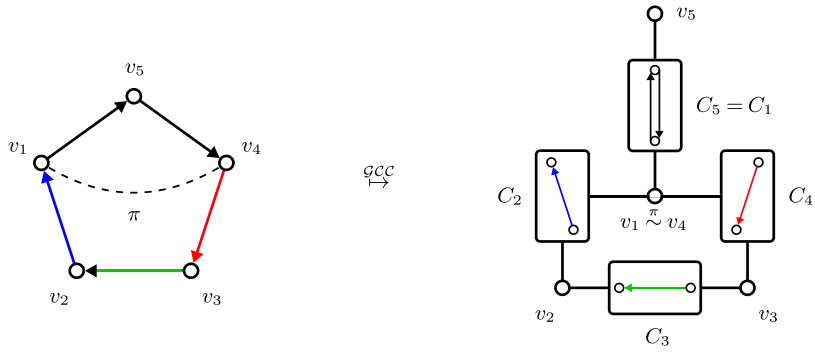


FIG. 6. *An example of $\mathcal{GCC}(T^\pi)$ in the case $\ell_n = \ell_1$ and $v_{n-1} \overset{\pi}{\sim} v_1$. For simplicity, we omit the trivial colored components consisting of a single isolated vertex.*

the directed path of $t_{N,i}$. Note that the alternating condition $\ell_1 \neq \ell_2 \neq \cdots \neq \ell_n$ implies that $C_{N,i} \neq C_{N,i-1}$ for each $i \geq 2$.

First, assume that $\ell_n \neq \ell_1$ so that

$$C_{N,1} \neq C_{N,2} \neq \cdots \neq C_{N,n} \neq C_{N,1}. \tag{14}$$

Note that the cycle c_N induces a cycle c_N^π in $\mathcal{GCC}(T_N^\pi)$, namely,

$$c_N^\pi = \big(\beta_\pi(v_n), C_{N,n}, \beta_\pi(v_{n-1}), C_{N,n-1}, \ldots, \beta_\pi(v_2), C_{N,2}, \beta_\pi(v_1), C_{N,1}, \beta_\pi(v_n)\big).$$

Since $\mathcal{GCC}(T_N^\pi)$ is a tree, c_N^π cannot be simple and must backtrack at some point, that is, there exists an $i \in [n]$ such that

$$C_{N,i} = C_{N,i-1} \quad \text{or} \quad \beta_\pi(v_i) = \beta_\pi(v_{i-1});$$

however, the first case contradicts (14), while the second case implies $v_i \sim_\pi v_{i-1}$.

If instead $\ell_n = \ell_1$, then $C_{N,n} = C_{N,1}$, and we obtain the cycle

$$c_N^\pi = \big(C_{N,n}, \beta_\pi(v_{n-1}), C_{N,n-1}, \ldots, \beta_\pi(v_2), C_{N,2}, \beta_\pi(v_1), C_{N,1}\big).$$

The only new case in the backtracking argument is

$$\beta_\pi(v_{n-1}) = \beta_\pi(v_1)$$

which would imply a shorter cycle

$$\big(\beta_\pi(v_{n-1}), C_{N,n-1}, \ldots, \beta_\pi(v_2), C_{N,2}, \beta_\pi(v_1)\big).$$

One can then proceed inductively. See Figure 6 for an illustration. □

Thus, it only remains to prove that $\lim_{N\to\infty} \tau_N^0[T_N^\pi] = 0$ for the test graph T_N in Lemma 4.8 and any partition $\pi \in \mathcal{P}(V)$ such that $\mathcal{GCC}(T_N^\pi)$ is not a tree.

4.3. *Proof of Proposition 2.4.* Carrying forward the notation from the previous section, we assume hereafter that the families $\mathbf{A}_{N,1}, \ldots, \mathbf{A}_{N,L}$ satisfy the assumptions of Proposition 2.4. In particular, assume that the families $\mathbf{A}_{N,2}, \ldots, \mathbf{A}_{N,L}$ are each permutation invariant. Without loss of generality, we can also assume that $\mathbf{A}_{N,1}$ is permutation invariant. Indeed, by Lemma 4.3 the quantity $\mathbb{E}[\frac{1}{N}\mathrm{Tr}(\varepsilon_N)]$ is a linear combination of terms of the form $\mathbb{E}[\frac{1}{N}\mathrm{Tr}[g(\mathbf{A}_N)]]$ for some graph monomials $g \in \mathcal{C}$. Let U_N be a uniform permutation matrix of order N independent of $\mathbf{A}_N$. By [18], Lemma 1.4, we know that

$$\mathbb{E}\left[\frac{1}{N}\mathrm{Tr}\big[g(\mathbf{A}_N)\big]\right] = \mathbb{E}\left[\frac{1}{N}\mathrm{Tr}\big[g\big(U_N \mathbf{A}_N U_N^*\big)\big]\right]$$

for any graph monomial g. Moreover, since the families $\mathbf{A}_{N,1},\dots,\mathbf{A}_{N,L},\{U_N\}$ are independent, we have the equality in (joint) distribution

$$\big(U_N\mathbf{A}_{N,1}U_N^*, U_N\mathbf{A}_{N,2}U_N^*,\dots,U_N\mathbf{A}_{N,L}U_N^*\big)\overset{d}{=}\big(U_N\mathbf{A}_{N,1}U_N^*,\mathbf{A}_{N,2},\dots,\mathbf{A}_{N,L}\big)$$

which proves that we can replace $\mathbf{A}_{N,1}$ with the permutation invariant family $U_N\mathbf{A}_{N,1}U_N^*$ in our calculations without consequence. We now prove the statement at the end of the previous section which will complete the proof of Proposition 2.4.

LEMMA 4.12. *For any partition $\pi\in\mathcal{P}(V)$ such that $\mathcal{GCC}(T_N^\pi)$ is not a tree, we have the asymptotic $\tau_N^0[T_N^\pi]=O(N^{-1})$.*

PROOF. Let ϕ be an arbitrary injective function from V^π to $[N]$. By [18], Lemma 2.18, the permutation invariance of $\mathbf{A}_N$ implies that

$$\tau_N^0[T_N^\pi]=\frac{1}{N}\frac{N!}{(N-|V^\pi|)!}\mathbb{E}\Big[\prod_{e\in E}A_{N,\eta_N(e)}^{(\gamma_N(e))}\big(\phi(e)\big)\Big]. \tag{15}$$

For any $\ell\in[L]$, let $T_{N,\ell}^\pi=(V_{N,\ell}^\pi,E_{N,\ell}^\pi)$ denote the test graph composed of the colored components of T_N^π in the color $\{\ell\}\times K$. The independence of the families $\mathbf{A}_{N,1},\dots,\mathbf{A}_{N,L}$ then implies that

$$\mathbb{E}\Big[\prod_{e\in E}A_{N,\eta_N(e)}^{(\gamma_N(e))}\big(\phi(e)\big)\Big]=\prod_{\ell=1}^{L}\mathbb{E}\Big[\prod_{e\in E_{N,\ell}^\pi}A_{N,\ell}^{(\gamma_N(e))}\big(\phi(e)\big)\Big].$$

Inverting the relationship (15) for each $T_{N,\ell}^\pi$, we obtain the expansion

$$\begin{aligned}\tau_N^0[T_N^\pi]&=\frac{1}{N}\frac{N!}{(N-|V^\pi|)!}\Bigg(\prod_{\ell=1}^{L}\frac{(N-|V_{N,\ell}^\pi|)!}{N!}\Bigg)N^{|\mathcal{CC}(T_N^\pi)|}\Bigg(\prod_{\ell=1}^{L}\tau_N^0[T_{N,\ell}^\pi]\Bigg)\\&=N^{-1+|V_\pi|-\sum_{\ell=1}^{L}|V_{N,\ell}^\pi|+|\mathcal{CC}(T_N^\pi)|}\big(1+O(N^{-1})\big)\prod_{\ell=1}^{L}\tau_N^0[T_{N,\ell}^\pi],\end{aligned} \tag{16}$$

where the asymptotic is uniform since $|V_{N,\ell}^\pi|\le|V^\pi|$ does not depend on N. Furthermore, recall that the graph of colored components $\mathcal{GCC}(T_N^\pi)=(\mathcal{V}_{N,\pi},\mathcal{E}_{N,\pi})$ satisfies $|\mathcal{V}_{N,\pi}|=|\mathcal{CC}(T_N^\pi)|+|V_\pi|$ and $|\mathcal{E}_{N,\pi}|=\sum_{\ell=1}^{L}|V_{N,\ell}^\pi|$. We can then rewrite (16) as

$$\tau_N^0[T_N^\pi]=N^{|\mathcal{V}_{N,\pi}|-1-|\mathcal{E}_{N,\pi}|}\big(1+O(N^{-1})\big)\prod_{\ell=1}^{L}\tau_N^0[T_{N,\ell}^\pi]. \tag{17}$$

We control the terms in the product above using the asymptotic (2) which requires the following definition.

DEFINITION 4.13. A *cut edge* of a finite graph is an edge whose deletion increases the number of connected components. We say that a finite connected graph is *two-edge connected* if it does not have any cut edges. Similarly, a *two-edge connected component* is a maximal two-edge connected subgraph.

For a finite graph G, we write $F(G)$ for its *forest of two-edge connected components*: the vertices of $F(G)$ are the two-edge connected components of G; the edges of $F(G)$ are the cut edges of G; and an edge connects the pair of two-edge connected components containing its adjacent vertices in G. In other words, one obtains $F(G)$ from G by contracting every two-edge connected component to a single vertex. One can easily verify that $F(G)$ is indeed

a forest. Finally, we define $\mathfrak{f}(G)$ to be the number of leaves in $F(G)$ with the convention that any trivial tree in $F(G)$ contributes two leaves to the count.

We can now prove an intermediate bound.

LEMMA 4.14. *The product in* (17) *satisfies the asymptotic*

$$\prod_{\ell=1}^{L} \tau_N^0[T_{N,\ell}^\pi] = O\big(N^{\sum_{\ell=1}^{L} \mathfrak{f}(T_{N,\ell}^\pi)/2 - |\mathcal{CC}(T_N^\pi)|}\big).$$

PROOF OF LEMMA 4.14. Let $\hat{T} = (\hat{V}, \hat{E})$ be a test graph. We use equation (11) to write

$$\tau_N^0[\hat{T}] = \sum_{\sigma \in \mathcal{P}(\hat{V})} N^{c(\hat{T}^\sigma) - c(\hat{T})} \tau_N[\hat{T}^\sigma] \mathrm{M}\,\ddot{\mathrm{o}}\mathrm{b}(0_V, \sigma).$$

The asymptotic (2) in the assumption of Proposition 2.4 implies that

$$\tau_N[\hat{T}^\sigma] = O\big(N^{\mathfrak{f}(\hat{T}^\sigma)/2 - c(\hat{T}^\sigma)}\big).$$

Furthermore, note that $\mathfrak{f}(\hat{T}^\sigma) \le \mathfrak{f}(\hat{T})$. Indeed, one can define the natural map $f_\sigma : V_{F(\hat{T})} \to V_{F(\hat{T}^\sigma)}$ from the two-edge connected components of $\hat{T}$ to the two-edge connected components of the quotient $\hat{T}^\sigma$ which is clearly surjective. Moreover, if C is a leaf in $F(\hat{T}^\sigma)$, then the fiber $f_\sigma^{-1}(C)$ necessarily contains a leaf in $F(\hat{T})$ which proves the inequality. Thus, we arrive at the asymptotics

$$\begin{aligned} \tau_N[\hat{T}^\sigma] &= O\big(N^{\mathfrak{f}(\hat{T})/2 - c(\hat{T}^\sigma)}\big); \\ \tau_N^0[\hat{T}] &= O\big(N^{\mathfrak{f}(\hat{T})/2 - c(\hat{T})}\big). \end{aligned} \tag{18}$$

Applying (18) to each of the test graphs $\hat{T} = T_{N,\ell}^\pi$, we obtain the stated result. In particular, there are only a finite number of possibilities for the underlying unlabeled undirected graph $\underline{T}_{N,\ell}^\pi$ of $T_{N,\ell}^\pi$ since the underlying unlabeled undirected graph $\underline{G}_N$ of T_N does not depend on N (see the paragraph before Lemma 4.8) which guarantees the uniformity of our asymptotic. □

Returning to our original task, we use Lemma 4.14 to rewrite (17) as

$$\tau_N^0[T_N^\pi] = O\big(N^{|\mathcal{V}_{N,\pi}| - 1 - |\mathcal{E}_{N,\pi}| + \sum_{\ell=1}^{L} \mathfrak{f}(T_{N,\ell}^\pi)/2 - |\mathcal{CC}(T_N^\pi)|}\big). \tag{19}$$

Recall that, for a finite undirected graph $G = (V, E)$, we have the identity

$$|E| - |V| = \sum_{v \in V} \left(\frac{\deg_G(v)}{2} - 1 \right).$$

We use this for our nontree $\mathcal{GCC}(T_N^\pi) = (\mathcal{V}_{N,\pi}, \mathcal{E}_{N,\pi})$ as follows. Let $\widetilde{\mathcal{G}}_{N,\pi} = (\widetilde{\mathcal{V}}_{N,\pi}, \widetilde{\mathcal{E}}_{N,\pi})$ be the graph obtained by *pruning* $\mathcal{GCC}(T_N^\pi)$: we remove the leaves of $\mathcal{GCC}(T_N^\pi)$ and their adjacent edges, iterating the procedure on the resulting graphs until no leaves remain. Since $\mathcal{GCC}(T_N^\pi)$ is not a tree, the pruned graph $\widetilde{\mathcal{G}}_{N,\pi}$ is nontrivial. For example, any simple cycle in $\mathcal{GCC}(T_N^\pi)$ will still remain in its entirety in $\widetilde{\mathcal{G}}_{N,\pi}$. Let $\widetilde{V}_{N,1}$ and $\widetilde{V}_{N,2}$ denote the vertices of

$\mathcal{CC}(T_N^\pi)$ and V^π, respectively, that remain in $\widetilde{\mathcal{G}}_{N,\pi}$. The exponent in (19) then becomes

$$
\begin{aligned}
&|\mathcal{V}_{N,\pi}| - 1 - |\mathcal{E}_{N,\pi}| + \sum_{\ell=1}^{L} \frac{\mathfrak{f}(T_{N,\ell}^\pi)}{2} - |\mathcal{CC}(T_N^\pi)| \\
&\quad = -1 - \big(|\widetilde{\mathcal{E}}_{N,\pi}| - |\widetilde{\mathcal{V}}_{N,\pi}|\big) + \sum_{C \in \mathcal{CC}(T_N^\pi)} \left(\frac{\mathfrak{f}(C)}{2} - 1\right) \\
&\quad = -1 - \sum_{C \in \widetilde{V}_{N,1}} \left(\frac{\deg_{\widetilde{\mathcal{G}}_{N,\pi}}(C)}{2} - 1\right) - \sum_{v \in \widetilde{V}_{N,2}} \left(\frac{\deg_{\widetilde{\mathcal{G}}_{N,\pi}}(v)}{2} - 1\right) \\
&\qquad + \sum_{C \in \mathcal{CC}(T_N^\pi)} \left(\frac{\mathfrak{f}(C)}{2} - 1\right).
\end{aligned} \tag{20}
$$

We claim that a colored component $C \in \mathcal{CC}(T_N^\pi)$ gets removed during the pruning procedure only if its forest of two-edge connected components $F(C)$ is the trivial tree. Indeed, for starters, note that $F(C)$ is always a tree since C is connected. If $F(C)$ is not the trivial tree, then it has at least two leaves $L_1 \neq L_2$. At the same time, we know that T_N^π is two-edge connected since it is the quotient of a two-edge connected graph T_N. This means that there are vertices v_1 and v_2 of the two-edge connected components L_1 and L_2, respectively, that are also connected by edges in T_N^π entirely outside of C. It follows that C belongs to a cycle in $\mathcal{GCC}(T_N^\pi)$ and hence does not get pruned. In particular,

$$\big\{C \in \mathcal{CC}(T_N^\pi) : \mathfrak{f}(C) > 2\big\} \subset \widetilde{V}_{N,1}.$$

Moreover, since one can follow this argument for any two leaves of $F(C)$, we immediately obtain the inequality $\mathfrak{f}(C) \leq \deg_{\widetilde{\mathcal{G}}_{N,\pi}}(C)$ for any $C \in \widetilde{V}_{N,1}$. Applying all of this to (20), we get

$$
\begin{aligned}
&|\mathcal{V}_{N,\pi}| - 1 - |\mathcal{E}_{N,\pi}| + \sum_{\ell=1}^{L} \frac{\mathfrak{f}(T_{N,\ell}^\pi)}{2} - \big|\mathcal{CC}(T_N^\pi)\big| \\
&\quad \leq -1 - \sum_{v \in \widetilde{V}_{N,2}} \left(\frac{\deg_{\widetilde{\mathcal{G}}_{N,\pi}}(v)}{2} - 1\right) \leq -1,
\end{aligned}
$$

where the last inequality follows by construction since $\deg_{\widetilde{\mathcal{G}}_{N,\pi}}(v) \geq 2$. Altogether, we conclude that

$$\tau_N^0[T_N^\pi] = O(N^{-1}). \qquad \square$$

4.4. *Proof of Theorem* 2.3. Theorem 6 in [21] ensures that

$$\operatorname{Tr}\big[g(\mathbf{A}_N)\big] \leq N^{\mathfrak{f}(g)/2} \prod_{e \in E} \big\| A_{N,\eta(e)}^{(\gamma(e))} \big\| \tag{21}$$

for any diagonal graph monomial g. The uniform operator norm bound on our matrices then implies the asymptotic (2) of Proposition 2.4, and the result follows.

4.5. *Proof of Theorem* 1.3. Let $\mathbf{A}_{N,1}, \ldots, \mathbf{A}_{N,L}$ and ε_N be as in Theorem 1.3. Without loss of generality, the LDC property allows us to restrict to the case of monomials $(P_{N,i})_{i \in [n]}$ in our matrix ε_N. By enlarging the index set K, if necessary, we write $\mathbf{A}_{N,L+1}$ for the family of coefficients $(D_{N,i,j})_{i \in [n],\, j \in \{0,\ldots,\deg(P_{N,i})\}}$ coming from the monomials $(P_{N,i})_{i \in [n]}$. We can then apply Theorem 2.3 to the families $\mathbf{A}_{N,1}, \ldots, \mathbf{A}_{N,L+1}$ and the monomials

$$\widetilde{P}_{N,i} = D_{N,i,0} X_{k_{N,i}(1)} D_{N,i,1} \cdots X_{k_{N,i}(d_{N,i})} D_{N,i,d_{N,i}},$$

where each $D_{N,i,j}$ can be written as $g_{N,i,j}(\mathbf{A}_N)$ for the graph monomial $g_{N,i,j}$ consisting of a single loop labeled by the matrix $D_{N,i,j} \in \mathbf{A}_{N,L+1} \subset \mathbf{A}_N = \bigsqcup_{\ell=1}^{L+1} \mathbf{A}_{N,\ell}$. Indeed, the LDC property implies that $\sup_{N\in\mathbb{N},i\in[n]} \deg(P_{N,i}) < \infty$. Since the full degree of $\widetilde{P}_{N,i}$ is bounded by $2\deg(P_{N,i}) + 1$, this guarantees the LFD property for the sequences $(\widetilde{P}_{N,i})_{N\in\mathbb{N},i\in[n]}$.

4.6. *Proof of Proposition 2.5.* We now consider the families $\widetilde{\mathbf{A}}_{N,1}, \ldots, \widetilde{\mathbf{A}}_{N,L}$, where

$$\widetilde{\mathbf{A}}_{N,\ell} = \big(A_{N,\ell}^{(k)} \circ \Gamma_{N,\ell}^{(k)}\big)_{k\in K}$$

and $\mathbf{\Gamma}_N = (\Gamma_{N,\ell}^{(k)})_{k\in K,\ell\in[L]}$ is a family of random matrices, independent of $\mathbf{A}_N$, satisfying the asymptotic (4).

We first prove that the conclusion of Theorem 2.3 holds for $(\widetilde{\mathbf{A}}_{N,\ell})_{\ell=1}^{L}$. We need only to prove the analogue of Lemma 4.12 for

$$\widetilde{\mathbf{A}}_N = \widetilde{\mathbf{A}}_{N,1} \sqcup \cdots \sqcup \widetilde{\mathbf{A}}_{N,L},$$

namely,

$$\tau_N^0\big[T_N^\pi(\widetilde{\mathbf{A}}_N)\big] = O\big(N^{-1}\big)$$

for any nontree $\mathcal{GCC}(T_N^\pi)$. Here, we use the notation $T_N^\pi(\widetilde{\mathbf{A}}_N)$ to indicate that the edge labels $\eta_N : E \to [L]$ and $\gamma_N : E \to K$ correspond to an assignment of matrices in the family $\widetilde{\mathbf{A}}_N$, as opposed to $\mathbf{A}_N$. The construction of the graph of colored components remains the same (in particular, the matrices $\mathbf{\Gamma}_N$ do not constitute another color). Recalling (15), the permutation invariance of $\mathbf{A}_N$ and the independence of $\mathbf{A}_N$ and $\mathbf{\Gamma}_N$ then imply that

$$\begin{aligned}
\tau_N^0\big[T_N^\pi(\widetilde{\mathbf{A}}_N)\big] &= \frac{1}{N} \sum_{\phi:V^\pi\hookrightarrow[N]} \mathbb{E}\bigg[\prod_{e\in E} A_{N,\eta_N(e)}^{(\gamma_N(e))}\big(\phi(e)\big)\Gamma_{N,\eta_N(e)}^{(\gamma_N(e))}\big(\phi(e)\big)\bigg] \\
&= \frac{1}{N} \sum_{\phi:V^\pi\hookrightarrow[N]} \bigg(\mathbb{E}\bigg[\prod_{e\in E} A_{N,\eta_N(e)}^{(\gamma_N(e))}\big(\phi(e)\big)\bigg]\mathbb{E}\bigg[\prod_{e\in E} \Gamma_{N,\eta_N(e)}^{(\gamma_N(e))}\big(\phi(e)\big)\bigg]\bigg) \\
&= \tau_N^0\big[T_N^\pi(\mathbf{A}_N)\big]\frac{(N-|V^\pi|)!}{N!} \sum_{\phi:V^\pi\hookrightarrow[N]} \mathbb{E}\bigg[\prod_{e\in E} \Gamma_{N,\eta_N(e)}^{(\gamma_N(e))}\big(\phi(e)\big)\bigg] \\
&= O\big(N^{-1}\big)O(1) = O\big(N^{-1}\big),
\end{aligned}$$

as was to be shown. Note that the last line follows from Lemma 4.12 and the assumption (4) on $\mathbf{\Gamma}_N$.

To finish, we prove the conclusion of Theorem 1.3 in the setting of Proposition 2.5. We reason as in Section 4.5. Let $\mathbf{A}_{N,L+1}$ denote the family of coefficients $D_{N,i,j}$ of the monomials $P_{N,i}$ defining ε_N. We would like to apply the previous method of proof to the families $\mathbf{A}_{N,1}, \ldots, \mathbf{A}_{N,L+1}$; however, the family $\mathbf{A}_{N,L+1}$ is not necessarily permutation invariant. To get around this, we define $\mathbf{\Gamma}_{N,L+1}$ to be the family of coefficients $D_{N,i,j}$ instead. We replace our previous definition of $\mathbf{A}_{N,L+1} = (A_{N,L+1}^{(k)})_{k\in K}$, where now $A_{N,L+1}^{(k)} = I_N$ is the identity matrix for every $k \in K$. Note that $\mathbf{A}_{N,L+1}$ is clearly permutation invariant, independent of $\mathbf{A}_{N,1} \sqcup \cdots \sqcup \mathbf{A}_{N,L}$ and uniformly bounded in operator norm.

We claim that the extended family $\mathbf{\Gamma}_N \sqcup \mathbf{\Gamma}_{N,L+1}$ still satisfies the asymptotic

$$\delta_N^0\big[T(\mathbf{\Gamma}_N \sqcup \mathbf{\Gamma}_{N,L+1})\big] := \mathbb{E}\bigg[\prod_{e\in E} \Gamma_{N,\eta(e)}^{(\gamma(e))}\big(\Phi(e)\big)\bigg] = O(1)$$

for any test graph T with edge labels $\eta : E \to [L+1]$ and $\gamma : E \to K$, where the asymptotic only depends on the underlying unlabeled undirected graph $\underline{T}$ of T. Indeed, for $E = E_1 \sqcup E_2 = \eta^{-1}([L]) \sqcup \eta^{-1}(L+1)$, we know that

$$\begin{aligned}|\delta_N^0[T(\mathbf{\Gamma}_N, \mathbf{\Gamma}_{N,L+1})]| &= \left|\mathbb{E}\left[\prod_{e\in E} \Gamma_{N,\eta(e)}^{(\gamma(e))}(\Phi(e))\right]\right| \\ &\le \mathbb{E}\left[\prod_{e\in E_1} |\Gamma_{N,\eta(e)}^{(\gamma(e))}(\Phi(e))| \prod_{e\in E_2} |\Gamma_{N,L+1}^{(\gamma(e))}(\Phi(e))|\right] \\ &= \mathbb{E}\left[\prod_{e\in E_1} |\Gamma_{N,\eta(e)}^{(\gamma(e))}(\Phi(e))|\right] O(1)\end{aligned}$$

since the coefficients $D_{N,i,j}$ satisfy a uniform operator norm bound. Note that the asymptotic is uniform since $|E_2| \le |E|$ is fixed by $\underline{T}$. Applying the Cauchy–Schwarz inequality, we can control the remaining term

$$\mathbb{E}\left[\prod_{e\in E_1} |\Gamma_{N,\eta(e)}^{(\gamma(e))}(\Phi(e))|\right] \le \mathbb{E}\left[\prod_{e\in E_1} |\Gamma_{N,\eta(e)}^{(\gamma(e))}(\Phi(e))|^2\right]^{\frac{1}{2}} = \delta_N^0[S(\mathbf{\Gamma}_N)]^{\frac{1}{2}},$$

where S is the test graph obtained from $T = (V, E, \mathrm{src}, \mathrm{tar}, \eta, \gamma)$ by adding an additional edge e' for every edge e in E in the opposite direction $\mathrm{src}(e') = \mathrm{tar}(e)$ and $\mathrm{tar}(e') = \mathrm{src}(e)$; the same $\eta(e') = \eta(e)$ label; and $\gamma(e')$ corresponding to the adjoint (recall that we assumed our families are closed under the adjoint $\cdot^*$):

$$\mathbf{A}_{N,\eta(e')}^{(\gamma(e'))} = (\mathbf{A}_{N,\eta(e)}^{(\gamma(e))})^*.$$

We use the asymptotic (4) for the test graph S to conclude that

$$\delta_N^0[T(\mathbf{\Gamma}_N, \mathbf{\Gamma}_{N,L+1})] \le \delta_N^0[S(\mathbf{\Gamma}_N)]^{\frac{1}{2}} O(1) = O(1),$$

as was to be shown. Again, the asymptotic is uniform since there are only finitely many possibilities for $\underline{S}$ given that $\underline{T}$ is fixed. Our work in the previous case then gives us the conclusion of Theorem 2.3 for the families $\widetilde{\mathbf{A}}_{N,1}, \ldots, \widetilde{\mathbf{A}}_{N,L+1}$ which proves the conclusion of Theorem 1.3 for the families $\widetilde{\mathbf{A}}_{N,1}, \ldots, \widetilde{\mathbf{A}}_{N,L}$.

Acknowledgments. The authors would like to thank the anonymous referees for their careful proofreading and helpful comments. This work was supported in part by the INSMI-CNRS through a PEPS JCJC grant. BA was supported in part by a Raymond H. Sciobereti Scholarship. GC was supported in part by the Simons Foundation and the Centre de Recherches Mathématiques through the Simons-CRM scholar-in-residence program. AD was supported in part by ERC Advanced Grant 669306. FG was supported in part by ERC Consolidator Grant 615897.

REFERENCES

[1] AJANKI, O. H., ERDŐS, L. and KRÜGER, T. (2017). Universality for general Wigner-type matrices. *Probab. Theory Related Fields* **169** 667–727. MR3719056 https://doi.org/10.1007/s00440-016-0740-2

[2] AU, B. (2018). Traffic distributions of random band matrices. *Electron. J. Probab.* **23** Paper No. 77, 48. MR3858905 https://doi.org/10.1214/18-EJP205

[3] BELINSCHI, S., DEMBO, A. and GUIONNET, A. (2009). Spectral measure of heavy tailed band and covariance random matrices. *Comm. Math. Phys.* **289** 1023–1055. MR2511659 https://doi.org/10.1007/s00220-009-0822-4

[4] BELINSCHI, S. T., MAI, T. and SPEICHER, R. (2017). Analytic subordination theory of operator-valued free additive convolution and the solution of a general random matrix problem. *J. Reine Angew. Math.* **732** 21–53. MR3717087 https://doi.org/10.1515/crelle-2014-0138

[5] BEN AROUS, G. and GUIONNET, A. (2008). The spectrum of heavy tailed random matrices. *Comm. Math. Phys.* **278** 715–751. MR2373441 https://doi.org/10.1007/s00220-007-0389-x

[6] BENAYCH-GEORGES, F., GUIONNET, A. and MALE, C. (2014). Central limit theorems for linear statistics of heavy tailed random matrices. *Comm. Math. Phys.* **329** 641–686. MR3210147 https://doi.org/10.1007/s00220-014-1975-3

[7] BENAYCH-GEORGES, F. and LÉVY, T. (2011). A continuous semigroup of notions of independence between the classical and the free one. *Ann. Probab.* **39** 904–938. MR2789579 https://doi.org/10.1214/10-AOP573

[8] BOEDIHARDJO, M. and DYKEMA, K. (2017). Asymptotic $*$-moments of some random Vandermonde matrices. *Adv. Math.* **318** 1–45. MR3689735 https://doi.org/10.1016/j.aim.2017.07.019

[9] BORDENAVE, C. and GUIONNET, A. (2013). Localization and delocalization of eigenvectors for heavy-tailed random matrices. *Probab. Theory Related Fields* **157** 885–953. MR3129806 https://doi.org/10.1007/s00440-012-0473-9

[10] CÉBRON, G., DAHLQVIST, A. and MALE, C. (2016+). Universal constructions for spaces of traffics. Preprint, https://arxiv.org/abs/1601.00168v1.

[11] CIZEAU, P. and BOUCHAUD, J.-P. (1994). Theory of Lévy matrices. *Phys. Rev. E* **50** 1810–1822.

[12] DESGROSEILLIERS, M., LÉVÊQUE, O. and MALE, C. (2015). Managing expectations: Freeness and the Fourier matrix. In *Twelfth International Symposium on Wireless Communication Systems*. Hal-01529534.

[13] GABRIEL, F. (2015+). Combinatorial theory of permutation-invariant random matrices I: Partitions, geometry and renormalization. Preprint, https://arxiv.org/abs/1503.02792v3.

[14] GABRIEL, F. (2015+). Combinatorial theory of permutation-invariant random matrices II: Cumulants, freeness and Levy processes. Preprint, https://arxiv.org/abs/1507.02465v2.

[15] GABRIEL, F. (2015+). A combinatorial theory of random matrices III: Random walks on $\mathfrak{S}(N)$, ramified coverings and the $\mathfrak{S}(\infty)$ Yang-Mills measure. Preprint, https://arxiv.org/abs/1510.01046v2.

[16] HELTON, J. W., MAI, T. and SPEICHER, R. (2018). Applications of realizations (aka linearizations) to free probability. *J. Funct. Anal.* **274** 1–79. MR3718048 https://doi.org/10.1016/j.jfa.2017.10.003

[17] KHORUNZHY, O., SHCHERBINA, M. and VENGEROVSKY, V. (2004). Eigenvalue distribution of large weighted random graphs. *J. Math. Phys.* **45** 1648–1672. MR2043849 https://doi.org/10.1063/1.1667610

[18] MALE, C. (2011+). Traffic distributions and independence: Permutation invariant random matrices and the three notions of independence. *Mem. Amer. Math. Soc.* To appear. Preprint, https://arxiv.org/abs/1111.4662v8.

[19] MALE, C. (2017). The limiting distributions of large heavy Wigner and arbitrary random matrices. *J. Funct. Anal.* **272** 1–46. MR3567500 https://doi.org/10.1016/j.jfa.2016.10.001

[20] MALE, C. and PÉCHÉ, S. (2014+). Uniform regular weighted graphs with large degree: Wigner's law, asymptotic freeness and graphons limit. Preprint, https://arxiv.org/abs/1410.8126v1.

[21] MINGO, J. A. and SPEICHER, R. (2012). Sharp bounds for sums associated to graphs of matrices. *J. Funct. Anal.* **262** 2272–2288. MR2876405 https://doi.org/10.1016/j.jfa.2011.12.010

[22] MINGO, J. A. and SPEICHER, R. (2017). *Free Probability and Random Matrices*. *Fields Institute Monographs* **35**. Springer, New York; Fields Institute for Research in Mathematical Sciences, Toronto, ON. MR3585560 https://doi.org/10.1007/978-1-4939-6942-5

[23] NICA, A. and SPEICHER, R. (2006). *Lectures on the Combinatorics of Free Probability*. *London Mathematical Society Lecture Note Series* **335**. Cambridge Univ. Press, Cambridge. MR2266879 https://doi.org/10.1017/CBO9780511735127

[24] SHLYAKHTENKO, D. (1996). Random Gaussian band matrices and freeness with amalgamation. *Int. Math. Res. Not.* **20** 1013–1025. MR1422374 https://doi.org/10.1155/S1073792896000633

[25] STANLEY, R. P. (2012). *Enumerative Combinatorics*. *Volume* 1, 2nd ed. *Cambridge Studies in Advanced Mathematics* **49**. Cambridge Univ. Press, Cambridge. MR2868112

[26] VOICULESCU, D. (1991). Limit laws for random matrices and free products. *Invent. Math.* **104** 201–220. MR1094052 https://doi.org/10.1007/BF01245072

[27] VOICULESCU, D. (1995). Operations on certain non-commutative operator-valued random variables. In *Recent Advances in Operator Algebras* (*Orléans*, 1992). *Astérisque* **232** 243–275. MR1372537

[28] VOICULESCU, D. V., DYKEMA, K. J. and NICA, A. (1992). *Free Random Variables*: *A Noncommutative Probability Approach to Free Products with Applications to Random Matrices*, *Operator Algebras and Harmonic Analysis on Free Groups*. *CRM Monograph Series* **1**. Amer. Math. Soc., Providence, RI. MR1217253

[29] ZAKHAREVICH, I. (2006). A generalization of Wigner's law. *Comm. Math. Phys.* **268** 403–414. MR2259200 https://doi.org/10.1007/s00220-006-0074-5