Singular value distribution of dense random matrices with block Markovian dependence

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Download date: 08. Feb. 2024

Singular value distribution of dense random matrices with block Markovian dependence

Jaron Sanders, Alexander Van Werde[*]

Eindhoven University of Technology, Department of Mathematics & Computer Science, The Netherlands

Abstract

A block Markov chain is a Markov chain whose state space can be partitioned into a finite number of clusters such that the transition probabilities only depend on the clusters. Block Markov chains thus serve as a model for Markov chains with communities. This paper establishes limiting laws for the singular value distributions of the empirical transition matrix and empirical frequency matrix associated to a sample path of the block Markov chain whenever the length of the sample path is $\Theta(n^2)$ with n the size of the state space.

The proof approach is split into two parts. First, we introduce a class of symmetric random matrices with dependent entries called approximately uncorrelated random matrices with variance profile. We establish their limiting eigenvalue distributions by means of the moment method. Second, we develop a coupling argument to show that this general-purpose result applies to the singular value distributions associated with the block Markov chain.

1. Introduction

Understanding hidden structures that underlie sequential data is an important challenge in data science. Not only do these structures give insight into the complex process which generates the data; once the structure is determined, any subsequent analysis can benefit from a reduction in dimensionality.

* Corresponding author.
  E-mail address: a.van.werde@tue.nl (A. Van Werde).

An attribute of sequential data is that past and future samples are dependent. Many tools for data analysis however have only been analyzed for independent data, because dependencies make the mathematical analysis of a system more difficult. Still, one can simply apply a tool originally designed for independent data and hope for insights. The conclusions one derives assuming independence will in the best scenario, but not necessarily, coincide with the conclusions that one would have derived in the appropriate dependent setting.

One such tool of the trade are spectral methods [1,7,70,72]. The term spectral methods refers here to any algorithm that makes use of the eigenvalues and eigenvectors of a matrix built from the data. Eigenvalues and eigenvectors are indeed routinely used to understand the underlying structure in data. If, for instance, the empirical eigenvalue distribution does not match the theoretical predictions associated with some model, then this may signify that the model is missing some important component of the process which generated the data [48,56,64,68]. Another example occurs in principal component analysis, where the eigenvalue distribution of an empirical covariance matrix can be used to detect the appropriate number of principal components [46,74]. If the matrix which is built from the data is non-Hermitian then algorithms typically employ singular values and singular vectors instead of eigenvalues and eigenvectors. Here recall that the ith largest singular value $s_i(M)$ of a square real matrix M is defined in terms of the ith largest eigenvalue of MM^{T} as $s_i(M) := (\lambda_i(MM^{\mathrm{T}}))^{1/2}$.

We are specifically interested in the singular value distribution of the *sample frequency matrix*

$$\hat{N}_X := (\hat{N}_{X,ij})_{i,j=1}^n \quad \text{where} \quad \hat{N}_{X,ij} := \sum_{t=0}^{\ell-1} \mathbb{1}_{X_t=i, X_{t+1}=j} \tag{1}$$

built from a dependent data sequence $X_0, X_1, \ldots, X_\ell$ taking values in $\{1, \ldots, n\}$ for some positive integer $n \in \mathbb{Z}_{\geq 1}$. The term *singular value distribution of* $\hat{N}_X$ here refers to the measure $\nu_{\hat{N}_X}$ defined by

$$\nu_{\hat{N}_X}([a,b]) := \frac{1}{n} \#\{i \in \{1, \ldots, n\} : a \leq s_i(\hat{N}_X) \leq b\} \tag{2}$$

for any $a < b$. While the singular value distribution of a random matrix is well understood when the elements are independent, this is not so when the matrix is constructed from dependent sequential data as is the case for $\hat{N}_X$.

This paper models the sequential data $X_0, X_1, \ldots, X_\ell$ as being generated by a block Markov chain. Block Markov chains are a model for dependent sequential data with an underlying community structure and have been used to develop and analyze community detection algorithms for sequential data; see [66,77]. Besides these two papers and the present paper, the only other rigorous analysis of the spectral properties of $\hat{N}_X$ when X is a block Markov chain can be found in [67]. There, an asymptotic distance between the K largest singular values and the $n - K$ smallest singular values is established.

The current paper establishes a limiting law for the singular value distribution of block Markovian random matrices such as $\hat{N}_X$ as the size n of the state space tends to infinity and the length of the data sequence satisfies $\ell = \Theta(n^2)$. For example, Theorem 1.1 describes the limiting law associated to $\hat{N}_X$, which is visualized in Fig. 2. Theorem 1.1 furthermore implies that the singular value distribution in block Markovian random matrices behaves as one would predict assuming the entries $\hat{N}_{X,ij}$ are independent. This gives some legitimacy to the practice of assuming some independence, particularly when the dependencies are asymptotically equivalent to those of a block Markov chain and the data sequence is sufficiently long. One can however
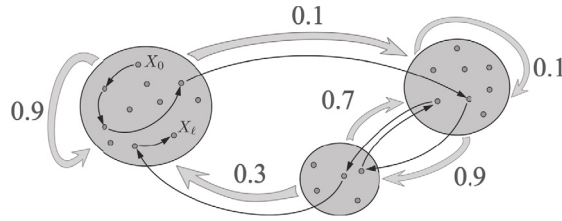
Fig. 1. Visualization of a block Markov chain on $K = 3$ clusters with $p = [[0.9, 0.1, 0], [0, 0.1, 0.9], [0.3, 0.7, 0]]$. The thick arrows visualize the cluster transition probabilities $p_{k,l}$ and the thin arrows visualise the transitions (X_t, X_{t+1}) of a sample path $(X_t)_{t=0}^{\ell}$. The starting point X_0 was chosen to lie in the leftmost cluster $\mathcal{V}_1$.

not ignore the Markovian dependence entirely since it causes the frequencies for different transitions to have different distributions. Indeed, the singular value distribution following from Theorem 1.1 does not necessarily agree with the singular value distribution which one would find assuming that the different entries $\hat{N}_{X,ij}$ are identically distributed.

We next state our main results in Section 1.1. This is followed by an overview of the literature in Section 2. Section 3 then provides notation and preliminaries in preparation of the proofs. Proof outlines are given in Section 4, and a brief comparison between our theoretical predictions and the singular value distribution obtained from an actual dataset is done in Section 5. Finally, all details for the proof are provided in Section 6.

1.1. Results

Our main object of study, which will subsequently be formally defined, are Markov chains which have a community structure. More precisely, the transition probabilities between states should only depend on the communities to which these states belong.

Fix a positive integer $K \in \mathbb{Z}_{\geq 1}$. For any $n \in \mathbb{Z}_{\geq K}$, pick a partition $(\mathcal{V}_k)_{k=1}^K$ of $\mathcal{V} := \{1, \ldots, n\}$ consisting only of nonempty sets. Let p be the transition matrix for an irreducible acyclic Markov chain on $\{1, \ldots, K\}$ with equilibrium distribution π. The *block Markov chain* with cluster transition matrix p and clusters $(\mathcal{V}_k)_{k=1}^K$ is then the Markov chain on $\mathcal{V}$ with transition probability $P_{i,j} := p_{x,y}/\#\mathcal{V}_y$ for every $i \in \mathcal{V}_x$ and $j \in \mathcal{V}_y$. Fig. 1 schematically depicts a block Markov chain.

In the subsequent results we are concerned with the asymptotic regime where n tends to infinity. Fix a sequence of strictly positive real numbers $\alpha := (\alpha_1, \ldots, \alpha_K)$ with $\sum_{k=1}^K \alpha_k = 1$ and assume that for every $k \in \{1, \ldots, K\}$ it holds that $\#\mathcal{V}_k = \alpha_k n + o(n)$. Let $X := (X_t)_{t=0}^{\ell}$ denote a sample path from the block Markov chain with an arbitrary starting distribution for X_0 and with length $\ell = \lambda n^2 + o(n^2)$ for some fixed $\lambda \in \mathbb{R}_{>0}$. Recall the definition of the empirical frequency matrix $\hat{N}_X$ from (1) and note that $\hat{N}_{X,ij}$ counts the number of traversals of edge (i, j). The *empirical transition matrix* $\hat{P}_X$ associated with the sample path is given by

$$\hat{P}_X := (\hat{P}_{X,ij})_{i,j=1}^n \quad \text{where} \quad \hat{P}_{X,ij} := \frac{\hat{N}_{X,ij}}{\sum_{k=1}^n \hat{N}_{X,ik}}. \tag{3}$$

It will be shown in Corollary 6.10 that there is no division by zero in (3) asymptotically almost surely.
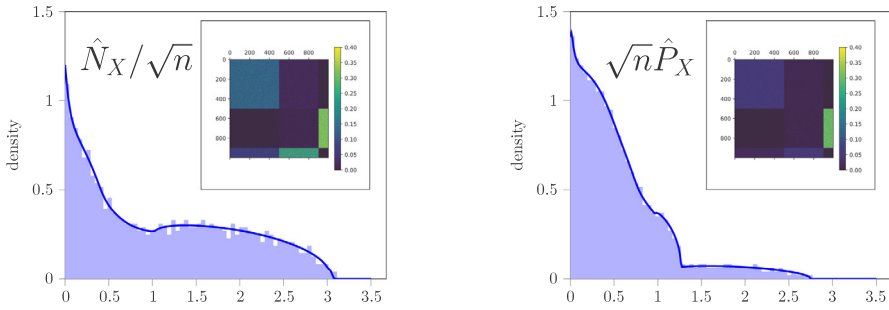
Fig. 2. On the left: $\hat{N}_X/\sqrt{n}$ and a frequency-based histogram of the singular values (bars) compared to our theoretical predictions (solid line). On the right: $\sqrt{n}\hat{P}_X$ and its singular values. These figures were made by sampling a path with $n = 1000$, $\ell = 2n^2$, $\alpha = (0.5, 0.4, 0.1)$ and cluster transition matrix $p = [[0.9, 0.1, 0], [0, 0.1, 0.9], [0.3, 0.7, 0]]$. The systems of equations in Theorems 1.1 and 1.2 were solved using the algorithm in [39, Proposition 4.1] after which we used the Stieltjes inversion formula (4) to recover the measures.

Some terminology is required to state the main results. The *Stieltjes transform* of a finite nonzero measure μ on $\mathbb{R}$ is the analytic function $s : \mathbb{C}^+ \to \mathbb{C}^-$ given by $s(z) = \int 1/(z - x)d\mu(x)$ where $\mathbb{C}^+ := \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$ is the upper half-plane and $\mathbb{C}^- := \{z \in \mathbb{C} : \mathrm{Im}(z) < 0\}$ is the lower half-plane. Let us remark that some authors refer to the Stieltjes transform by another name such as *Cauchy transform* or *Cauchy–Stieltjes transform*. Further, let us warn that some authors employ a convention which differs by a minus sign from the notation employed here; they instead consider the map $z \mapsto \int 1/(x - z)d\mu(x)$. The relevance of the Stieltjes transform for our purposes is that μ can be recovered from $s(z)$ by the Stieltjes inversion formula [8, Theorem B.8.] which states that for any continuity points $a < b$

$$\mu([a, b]) = -\frac{1}{\pi} \lim_{\varepsilon \to 0^+} \int_a^b \mathrm{Im}(s(x + \sqrt{-1}\varepsilon))dx. \tag{4}$$

A sequence of random measures μ_n on $\mathbb{R}$ is said to *converge weakly in probability* to a finite measure μ if $\int f(x)d\mu_n(x) \to \int f(x)d\mu(x)$ in probability for every continuous bounded function $f \in \mathcal{C}_b(\mathbb{R})$. Finally, the *symmetrization* of a measure μ on $\mathbb{R}_{\geq 0}$ is the measure $\mathrm{sym}(\mu)$ on $\mathbb{R}$ given by $A \mapsto (\mu(A \cap \mathbb{R}_{\geq 0}) + \mu((-A) \cap \mathbb{R}_{\geq 0}))/2$ where A ranges over all measurable subsets of $\mathbb{R}$ and $-A := \{-a : a \in A\}$.

Visualizations of the following results are shown in Fig. 2.

Theorem 1.1. *The empirical singular value distribution $\nu_{\hat{N}_X/\sqrt{n}}$ converges weakly in probability to a compactly supported probability measure ν on $\mathbb{R}_{\geq 0}$. Moreover, the symmetrization $\mathrm{sym}(\nu)$ has Stieltjes transform $s(z) = \sum_{i=1}^K \alpha_i(a_i(z) + a_{K+i}(z))/2$ where $a_1, \ldots, a_{2K}$ are the unique analytic functions from $\mathbb{C}^+$ to $\mathbb{C}^-$ such that the following system of equations is satisfied*

$$a_i(z)^{-1} = z - \sum_{j=1}^K \lambda\pi(i)\alpha_i^{-1} p_{i,j} a_{K+j}(z) \tag{5}$$

$$a_{i+K}(z)^{-1} = z - \sum_{j=1}^K \lambda\pi(j)\alpha_i^{-1} p_{j,i} a_j(z) \tag{6}$$

for $i = 1, \ldots, K$.

Theorem 1.2. *The empirical singular value distribution $v_{\sqrt{n}\hat{P}_X}$ converges weakly in probability to a compactly supported probability measure v on $\mathbb{R}_{\geq 0}$. Moreover, the symmetrization $\mathrm{sym}(v)$ has Stieltjes transform $s(z) = \sum_{i=1}^{K} \alpha_i(a_i(z) + a_{K+i}(z))/2$ where $a_1, \ldots, a_{2K}$ are the unique analytic functions from $\mathbb{C}^+$ to $\mathbb{C}^-$ such that the following system of equations is satisfied*

$$a_i(z)^{-1} = z - \sum_{j=1}^{K} \lambda^{-1} \pi(i)^{-1} \alpha_i p_{i,j} a_{K+j}(z) \tag{7}$$

$$a_{i+K}(z)^{-1} = z - \sum_{j=1}^{K} \lambda^{-1} \pi(j)^{-1} \alpha_i^{-1} \alpha_j^2 p_{j,i} a_j(z) \tag{8}$$

for $i = 1, \ldots, K$.

Observe that the limiting law described in Theorem 1.1 is the same as occurs for a random matrix M with mean-zero independent entries matching the variance profile of $\hat{N}_X$ [79, Theorem 6.1]. By matching variance profile, it is here meant that $\mathrm{Var}[M_{ij}] = \mathrm{Var}[\hat{N}_{X,ij}]$ for all i, j. Similarly, the limiting law in Theorem 1.2 corresponds to the limiting law of a random matrix with independent entries which instead matches the variance profile of $\sqrt{n}\hat{P}_X$. This hints at an underlying more general universality principle, which is commonplace in random matrix theory. Informally, universality states that the spectrum of a random matrix often only depends on the variance of the entries.

Indeed, the first step of our proof establishes a precise version of this universality statement in Corollary 4.3. The proof strategy is essentially a modification of the moment method. More specifically, we generalize a result from [40] concerning the eigenvalue distributions of approximately uncorrelated random matrices to include the possibility of a variance profile. The explicit description of the Stieltjes transform of the limiting laws relies on [79].

The second step of our proof establishes that this general-purpose universality principle applies to the block Markovian random matrices $\hat{N}_X$ and $\hat{P}_X$. Proposition 4.7 contains the corresponding result. The key difficulty is to control the dependence. To this end, we use a coupling-based approach which is original to this paper and in fact the main new ingredient to establish the results. The reader is referred to Proposition 4.8 for a special case of Proposition 4.7 whose proof contains the key ideas.

For future research, an investigation into what happens when the sample path is much shorter, i.e., $\ell = o(n^2)$, can be considered. We anticipate that the results of this paper can be extended to such regimes, in which the empirical frequency matrix is sparser. Nontrivial modifications would however be required in the part of the proof which relies on the moment method. This is because, in a sparse random matrix, normalizing for the variance causes all higher moments of the entries to diverge. This issue can already be observed at the level of a scalar random variable. Namely consider a sequence of Bernoulli random variables ξ_n with probability of success $p_n = o(1)$ as n tends to infinity, and set $\zeta_n := (\xi_n - p_n)/\sqrt{p_n(1 - p_n)}$. Then $\mathrm{Var}[\zeta_n] = 1$ whereas $\mathbb{E}[\zeta_n^3] = \Theta(p_n^{-1/2})$ diverges.

2. Related literature

2.1. Block Markov chains

Block Markov chains are the Markov chain analogue of the stochastic block model, and can similarly be used as a benchmark to investigate community detection problems. Community detection has been studied extensively within the context of the stochastic block model, and

we refer interested readers to [35] for an overview. Community detection problems within the context of block Markov chains received attention more recently [30,31,66,77,78,80].

Spectral clustering algorithms for learning low-rank structures in Markov chains from trajectories that specifically utilize the sample frequency matrix $\hat{N}_X$ or the empirical transition matrix $\hat{P}_X$ have been analyzed in [66,77]. The fact that $\hat{N}_X$ and $\hat{P}_X$ have previously been used in the context of community detection algorithms for Markov chains is also what motivated us to consider these two specific matrices. One could in principle however also build different matrices from the data. Appropriately modified, the methods of the current paper should still apply and thereby allow one to derive the associated singular value distributions.

In order to compare algorithmic performance to an information-theoretical lower bound on the detection error rate satisfied under any clustering algorithm, [66] required a sufficiently sharp upper bound to the largest singular value of $\hat{N}_X - \mathbb{E}[\hat{N}_X]$. The singular values of $\hat{N}_X$ were recently also considered in [67]. It is established there that $\hat{N}_X$ has K informative singular values of size $\Theta(\ell/n)$, and that the remaining $n - K$ singular values are $O(\sqrt{\ell/n})$. Besides the dense regime $\ell = \Omega(n^2)$, the sparser regimes $\ell = \Omega(n \ln n)$ and $\ell = \omega(n)$ are also considered in [66,67].

2.2. Random matrices generated by stochastic processes

The spectral distributions of matrices whose entries are sampled by means of a stochastic process, such as a Markov chain, were considered in [22,34,53,54,60,62]. These results are similar to ours in that the randomness is due to the sampling noise of the stochastic process, but differ in the precise construction of the matrix.

Sample covariance matrices for time series have been considered in [9,10,44,55,61,75]. Let us note that covariance matrices of time series can also be viewed as an instance of the aforementioned study of random matrices with entries sampled from a stochastic process by consideration of the entries of the data matrix. Sample autocovariance matrices of time series have been considered in [12,17,18,49,50,76].

2.3. Coupling arguments

The critical new ingredient in our proof are the coupling arguments which are used to establish Proposition 4.7. Coupling arguments are a natural way to deal with the dependence in a Markov chain. They have been used in this setting since the seminal paper [29]. Coupling arguments in random matrix theory are however not common place; exceptions we are aware of are [9,11].

2.4. Approximately uncorrelated random matrices

A class of self-adjoint random matrices with dependent entries and decaying covariances, called approximately uncorrelated random matrices, was studied in [40]. The authors establish that the empirical eigenvalue distribution of an approximately uncorrelated random matrix converges weakly in probability to the semicircular law. Our Corollary 4.3 generalizes their approach to admit the possibility that not all entries have the same variance. Improved results to convergence weakly almost surely are established in [19,33] under additional assumptions. It would be interesting to establish similar results in the presence of variance profiles (see the remarks preceding Proposition 6.16).

2.5. Random matrices with a variance profile

Let M be a random matrix with independent centered entries and variance profile $S_{ij} :=$ Var$[M_{ij}]$. The classical results on the spectral properties of random matrices assume that this variance profile is constant, but extensions to nonconstant variance profiles have also been considered [2,4,27,28,32,37].

It is typically necessary to assume that the variance profile has tractable asymptotic behavior. One notion of tractable asymptotic behavior may be found by employing notions from graphon theory [23,26,79]. In this case the variance profile converges to an integrable function $W : [0, 1]^2 \rightarrow \mathbb{R}$. Let us note that results characterizing eigenvalue distributions in this setting historically preceded the graphon-theoretic terminology; see [71]. Graphon theory was originally developed in [52] as limiting objects for sequences of dense graphs.

Systems of self-consistent equations as in Theorem 1.1 frequently occur in the theory of random matrices with variance profiles [32,71,79]. However, solving these equations to determine an explicit expression for the Stieltjes transform $s(z)$ is rarely possible. A numerical method based on an iteration of contraction maps has been developed in the field of operator-valued free probability theory [39].

2.6. Poisson limit theorems for Markov chains

The variance profiles in our block Markovian setting follow from a Poisson limit theorem; see Theorem 4.5. This is in turn deduced from a nonasymptotic Poisson approximation theorem; see Theorem 6.11. Poisson limit theorems for Markov chains are a topic of study in their own right. We refer to [69, Section 2.3] and [38, Section 5] for an overview of the literature.

Distinct from the literature, the emphasis of Theorem 6.11 lies on the fact that the state space is growing. Compare this e.g. to [63] which concerns the number of visits to an increasingly rare cylindrical set in the sample path of a Markov chain on a *fixed* finite state space. Theorem 6.11's proof relies on a general Poisson approximation result for sums of dependent random variables from [6] which in turn relies on a method from [24].

2.7. Random transition matrices

There has been recent interest in the spectral properties of random walks in a random environment. The setting of [15] is to first sample a random $n \times n$ matrix M of independent and identically distributed nonnegative real random variables of finite variance, and to then construct the random transition matrix $P := D^{-1}M$ with D the diagonal matrix containing the row sums of M. The results then concern limiting laws for the singular value distribution and eigenvalue distribution of $\sqrt{n}P$. Different models for random transition matrices have been considered in [13,14,16,20,21,25,42,43,47,56,58,81].

Our study of $\hat{P}_X$ differs from the study of P in the source of randomness: the randomness in $\hat{P}_X$ is due to the observation noise in the sampled Markov chain in a deterministic environment whereas the randomness in P is due to a perfect observation of a random environment. Our situation involves the additional subtlety that one has to deal with the dependence intrinsic to the sampling noise of a Markov chain. It should however be mentioned that many of the aforementioned results also concern the distribution of eigenvalues, while we study singular values.

3. Notation and preliminaries

3.1. Block Markov chains

Block Markov chains were defined in Section 1.1. Denote $\sigma : \mathcal{V} \to \{1, \dots, K\}$ for the map which sends any $v \in \mathcal{V}_k$ to the cluster index k. Note that the definition of a block Markov chain implies that $\Sigma_X := (\sigma(X_t))_{t=0}^\ell$ is a Markov chain on the space of clusters $\{1, \dots, K\}$ with transition matrix p.

Recall that it is assumed that $|\#\mathcal{V}_k - \alpha_k n| = o(n)$ with $\alpha_k > 0$ for all $k \in \{1, \dots, K\}$. Further, recall that the $\mathcal{V}_k$ are assumed to be nonempty for all n. Hence, there exists some $\alpha_{min} \in \mathbb{R}_{>0}$, independent of n, such that $\#\mathcal{V}_k > \alpha_{min} n$ for all $k \in \{1, \dots, K\}$.

We denote $E_X := (E_{X,t})_{t=1}^\ell$ for the chain of edges $E_{X,t} := (X_{t-1}, X_t)$ associated with X.

3.2. Graphs

All graphs in this paper are assumed to be finite and are allowed to have self-loops or multiple edges. We use the term *simple* to refer to the case where self-loops or multiple edges are not allowed. An *ordered tree* is a simple rooted tree such that every vertex is equipped with a total order on its descendants. The collection of all ordered trees on $k + 1$ vertices is denoted by $\mathcal{T}_k$.

For any $n \in \mathbb{Z}_{\geq 1}$ we denote $\vec{E}_n$ for the set of directed edges $\{1, \dots, n\}^2$ corresponding to the state space $\mathcal{V} := \{1, \dots, n\}$. Given a directed edge $e = (i, j)$ and an $n \times n$ matrix M we denote $M_e := M_{ij}$. For two vectors of integers $m, m' \in \mathbb{Z}^R$ we denote $m \leq m'$ if $m_i \leq m_i'$ for all $i \in \{1, \dots, R\}$.

3.3. Graphon theory

A *graphon* is an integrable map $W : [0, 1]^2 \to \mathbb{R}$ which is symmetric, meaning that $W(x, y) = W(y, x)$ for all $x, y \in [0, 1]$. We denote $\mathcal{W}_0$ for the collection of graphons W such that $0 \leq W(x, y) \leq 1$ for all $x, y \in [0, 1]$.

The *cut norm* on a graphon W is defined by $\|W\|_\square := \sup_{S,T \subseteq [0,1]} |\int_{S \times T} W(x, y) \mathrm{d}x \mathrm{d}y|$ where the supremum runs over all measurable subsets S, T of $[0, 1]$. The *cut metric* $\delta_\square$ on the space of graphons is defined by

$$\delta_\square(W, W') := \inf_\phi \|W^\phi - W'\|_\square \tag{9}$$

where the infimum runs over all measure preserving bijections $\phi : [0, 1] \to [0, 1]$ and $W^\phi(x, y) := W(\phi(x), \phi(y))$.

Given a symmetric matrix $M \in \mathbb{R}^{n \times n}$ one can define a graphon W^M by setting

$$W^M(x, y) := M_{ij} \quad \text{if } (x, y) \in \left[\frac{i-1}{n}, \frac{i}{n}\right) \times \left[\frac{j-1}{n}, \frac{j}{n}\right). \tag{10}$$

The graphon W^M can be assigned values on the boundaries $x = 1$ and $y = 1$ by extending continuously.

3.4. Measure theory

Recall that the *Stieltjes transform* of a finite measure on $\mathbb{R}$, *weak convergence in probability of random measures*, and the *symmetrization* of a measure on $\mathbb{R}_{\geq 0}$ were defined in Section 1.1.

For two probability measures μ and ν on a countable space $\mathcal{V}$, we can define the *total variation distance* as

$$d_{\mathrm{TV}}(\mu, \nu) := \frac{1}{2} \sum_{x \in \mathcal{V}} |\mu(x) - \nu(x)|. \tag{11}$$

Whenever the kth moment of a measure μ on $\mathbb{R}$ exists, $k \in \mathbb{Z}_{\geq 0}$, the kth moment will be denoted as $\mathfrak{m}_k(\mu) := \int x^k d\mu(x)$. Note that the definition of the *empirical singular value distribution* of an $n \times n$ real matrix M with singular values $s_1(M) \geq \cdots \geq s_n(M)$ from (2) can be rephrased as stating that ν_M is the measure on $\mathbb{R}_{\geq 0}$ given by $\nu_M := n^{-1} \sum_{i=1}^{n} \delta_{s_i(M)}$ where $\delta_{s_i(M)}$ denotes a point mass at $s_i(M)$. The *empirical eigenvalue distribution* μ_A of an $n \times n$ symmetric matrix A with eigenvalues $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$ is similarly defined as the measure on $\mathbb{R}$ given by $\mu_A := n^{-1} \sum_{i=1}^{n} \delta_{\lambda_i(A)}$.

Denote the *Hermitian dilation of M* by

$$H(M) := \begin{pmatrix} 0 & M \\ M^{\mathrm{T}} & 0 \end{pmatrix}. \tag{12}$$

Note that the eigenvalues of $H(M)$ are $s_1(M), -s_1(M), \ldots, s_n(M), -s_n(M)$. Hence, $\mu_{H(M)} = \mathrm{sym}(\nu_M)$ where $\mathrm{sym}(\nu_M)$ denotes the symmetrization of the measure ν_M.

3.5. Compressed notation for conditional probability

Let $A, B, C, D \in \mathcal{F}$ be events in some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{P}(C \cap D) \neq 0$. We will on a few occasions encounter long expressions involving the associated conditional probability. To preserve readability we may then also use the following compressed notation

$$\mathbb{P}(A, B \mid C, D) =: \mathbb{P}\begin{pmatrix} A & \bigm| & C \\ B & \bigm| & D \end{pmatrix}. \tag{13}$$

Similar notation may be used for conditional expectation and unconditional probability.

3.6. Asymptotic notation

We employ the usual conventions for big-O notation: Let $(x_n)_{n=0}^{\infty}$ and $(y_n)_{n=0}^{\infty}$ be two sequences of real numbers. Then $x_n = O(y_n)$ if and only if there exist $C, n_0 > 0$ such that $|x_n| \leq C|y_n|$ for all $n \geq n_0$. Similarly, $x_n = o(y_n)$ if and only if for every $C > 0$ there exists some n_0 such that $|x_n| \leq C|y_n|$ for all $n \geq n_0$; and $x_n = \Omega(y_n)$ if and only if there exist $C, n_0 > 0$ such that $|x_n| \geq C|y_n|$ for all $n \geq n_0$. Finally, $x_n = \Theta(y_n)$ if and only if $x_n = O(y_n)$ as well as $x_n = \Omega(y_n)$.

If $(x_n)_{n=0}^{\infty}$ depends on some parameters a, b, then the possible dependence of the constants on the parameters is expressed in the notation. For example, $x_n = O_a(y_n)$ means that there exist $C, n_0 > 0$ possibly dependent on a but not on b such that $|x_n| \leq C|y_n|$ for all $n \geq n_0$. When emphasizing like such, the parameters a, b are assumed not to depend on n.

4. Proof outline

Our proof of Theorems 1.1 and 1.2 consists of two parts: a general-purpose universality result and a reduction argument.

The first part of our proof is given in Section 4.1 where we generalize the results concerning eigenvalues of approximately uncorrelated random matrices from [40] to admit the possibility

of a variance profile. The corresponding general-purpose universality results are Theorem 4.2 and Corollary 4.3. Similar results for matrices with variance profile have previously also appeared in [79, Theorem 3.2 and Theorem 3.4] under the assumption that the entries are independent. The combination of variance profiles with dependence, i.e., the notion of approximately uncorrelated random matrices with a variance profile, is new.

The second part of our proof is given in Section 4.2 and consists of a reduction to Corollary 4.3. This involves two key difficulties. First, we need to determine the variance profiles associated with the block Markovian random matrices $\hat{N}_X$ and $\hat{P}_X$. These variance profiles are established using Theorem 4.5, which states that $\hat{N}_{X,ij}$ is asymptotically Poisson distributed with a rate that depends only on the clusters to which i and j belong. Second, we need to establish that the block Markovian random matrices are in the approximately uncorrelated universality regime. To do so, we develop a coupling argument which shows that the covariance between the number of traversals of different edges decays sufficiently quickly; see Proposition 4.7 for the corresponding result.

4.1. Eigenvalue distributions of approximately uncorrelated random matrices with variance profile

The results in this section concern the eigenvalues of a symmetric matrix whereas we are interested in the singular values of the matrices $\hat{N}_X$ and $\hat{P}_X$. To this end, let us remind the reader of the fact that the study of singular values of any matrix can be reduced to the study of the eigenvalues of a symmetric matrix by a Hermitian dilation (recall (12)).

Definition 4.1. A family of symmetric random matrices $(A_n)_{n=1}^{\infty}$ is said to be *approximately uncorrelated with variance profile* if, for any non-negative integers $0 \leq r \leq R$ and $m_1, \ldots, m_R \in \mathbb{Z}_{\geq 0}$ with $m_i = 1$ for $i = 1, \ldots, r$, it holds that

$$\max_{\forall k \neq l: \{i_k, j_k\} \neq \{i_l, j_l\}} \left| \mathbb{E} \left[A_{n, i_1 j_1}^{m_1} \cdots A_{n, i_R j_R}^{m_R} \right] \right| = O_{m,R}(n^{-r/2}) \tag{14}$$

and

$$\max_{\forall k \neq l: \{i_k, j_k\} \neq \{i_l, j_l\}} \left| \mathbb{E} \left[A_{n, i_1 j_1}^2 \cdots A_{n, i_R j_R}^2 \right] - \prod_{k=1}^{R} \mathbb{E} \left[A_{n, i_k j_k}^2 \right] \right| = o_R(1) \tag{15}$$

where the maxima run over all values of $(i_1, j_1), \ldots, (i_R, j_R) \in \{1, \ldots, n\}^2$ with $\{i_k, j_k\} \neq \{i_l, j_l\}$ for all $k \neq l$.

In order to identify a limit of the empirical eigenvalue distribution $\mu_{A_n/\sqrt{n}}$ it is necessary to assume that the variance profile has tractable asymptotic behavior. We follow the approach taken in [79, Theorem 3.2] and employ the homomorphism density. The *homomorphism density* from a simple graph $F = (V, E)$ on $V = \{1, \ldots, R\}$ to a symmetric matrix $M \in \mathbb{R}^{n \times n}$ is defined by

$$t(F, M) := \frac{1}{n^R} \sum_{i_1, \ldots, i_R = 1}^{n} \prod_{\{v, w\} \in E} M_{i_v, i_w}. \tag{16}$$

The name homomorphism density may be explained by the fact that if A is the adjacency matrix of a graph G, then $t(F, A)$ counts the number of graph homomorphisms of F to G. A detailed proof for the following result may be found in Section 6.1.

Theorem 4.2. *Let $(A_n)_{n=1}^{\infty}$ be a family of symmetric random matrices which are approximately uncorrelated with variance profile $S_n := (\mathrm{Var}[A_{n,ij}])_{i,j=1}^{n}$. Assume that for every ordered tree $T \in \cup_{m=0}^{\infty} \mathcal{T}_m$ it holds that $t(T, S_n)$ has a limit as $n \to \infty$. Then, the empirical eigenvalue distribution of $\mu_{A_n/\sqrt{n}}$ converges weakly in probability to the unique probability measure μ whose moments are given by*

$$\mathfrak{m}_{2m}(\mu) = \sum_{T \in \mathcal{T}_m} \lim_{n \to \infty} t(T, S_n), \quad \mathfrak{m}_{2m+1}(\mu) = 0, \tag{17}$$

for every $m \in \mathbb{Z}_{\geq 0}$. Moreover, μ is compactly supported.

Proof sketch. Just as in the classical moment method, the key step is to show that $\mathbb{E}[\mathfrak{m}_k(\mu_{A_n/\sqrt{n}})] = \mathfrak{m}_k(\mu) + o_k(1)$ for every $k \in \mathbb{Z}_{\geq 0}$. Observe that

$$\mathbb{E}[\mathfrak{m}_k(\mu_{A_n/\sqrt{n}})] = n^{-1-k/2}\mathbb{E}[\mathrm{Tr}(A_n^k)] \tag{18}$$

$$= n^{-1-k/2} \sum_{i_1,\ldots,i_k=1}^{n} \mathbb{E}[A_{n,i_1i_2} A_{n,i_2i_3} \cdots A_{n,i_{k-1}i_k} A_{n,i_ki_1}] \tag{19}$$

for every $k \in \mathbb{Z}_{\geq 0}$. Given a sequence of indices $i := (i_1, \ldots, i_k, i_1)$, which occurs on the right-hand side of (19), let $G_i := (V(i), E(i))$ denote the induced undirected graph with vertex set $V(i) := \{i_1, \ldots, i_k\}$ and edge set $E(i) := \{\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_k, i_1\}\}$. Viewing i as a cycle on G_i, let $r_1(i)$ be the number of edges which are traversed exactly once and let $r_2(i)$ be the number of edges which are traversed exactly twice. Note that we could in principle also define $r_3(i), r_4(i), \ldots$, but these quantities will not be relevant in the proof.

Consider $P(i) := \mathbb{E}[A_{n,i_1i_2} \cdots A_{n,i_ki_1}]$. In a classical application of the moment method, one has assumed that all entries $A_{n,ij}$ with $i \leq j$ are independent and centered. Under such assumptions, it immediately follows that $P(i) = 0$ whenever $r_1(i) > 0$. The entries of A_n are however not independent in our case. It may thus be that $P(i) \neq 0$. Instead, one has to rely on part (14) in the definition of an approximately uncorrelated random matrix to deduce that $P(i)$ is small whenever $r_1(i)$ is large. Combined with a bound on the number of terms with $r_1(i) = r$, which is stated in Lemma 6.4 and established in [40], this is still sufficient to argue that the contribution of the terms with $r_1(i) > 0$ is asymptotically negligible.

When $k = 2m + 1$ is odd, the number of terms with $r_1(i) = 0$ in (19) is of a smaller order than the normalizing factor $n^{-1-k/2}$. This yields that $n^{-1-k/2}\mathbb{E}[\mathrm{Tr}(A_n^k)] = o_k(1)$ for all odd values of k. When $k = 2m$ is even, the asymptotics are dominated by the contribution of those $P(i)$ for which G_i is a tree and $r_2(i) = k/2$. This leads to the conclusion that $n^{-1-k/2}\mathbb{E}[\mathrm{Tr}(A_n^k)] = \sum_{T \in \mathcal{T}_m} t(T, S_n) + o_k(1)$. $\square$

Note that Theorem 4.2 also applies to random matrices with independent entries. Correspondingly, the limit μ may be explicitly identified whenever it is known for independent random matrices with the same variance profile. The following corollary is an instance of this principle and uses the description provided in [79, Theorem 3.4] for eigenvalues of matrices with independent entries. The details for the reduction argument are provided in Section 6.2.

Corollary 4.3. *Let $(A_n)_{n=1}^{\infty}$ be a family of symmetric random matrices which are approximately uncorrelated with variance profile $S_n := (\mathrm{Var}[A_{n,ij}])_{i,j=1}^{n}$. Assume that there exists some graphon $W \in \mathcal{W}_0$ such that $\delta_{\square}(W^{S_n}, W) \to 0$. Then, the empirical eigenvalue distribution $\mu_{A_n/\sqrt{n}}$ converges weakly in probability to the probability measure μ whose Stieltjes transform $s(z)$ is given by*

$$s(z) = \int_0^1 a(z, x)dx \tag{20}$$

where $a(z, x)$ is the unique analytic function from $\mathbb{C}^+ \times [0, 1]$ to $\mathbb{C}^-$ satisfying the following self-consistent equation

$$a(z, x)^{-1} = z - \int_0^1 W(x, y)a(z, y)dy. \tag{21}$$

4.2. Block Markovian random matrices are approximately uncorrelated with variance profile

The proof of Theorems 1.1 and 1.2 now amounts to a reduction to Corollary 4.3 which is done in three steps. First, in Section 4.2.1 we argue that we may recenter the matrices and we may assume that X starts from its equilibrium distribution. Second, in Section 4.2.2 we determine the variance profiles by means of a Poisson limit theorem. Finally, in Section 4.2.3 we show that the random matrices are in the approximately uncorrelated regime which is done using a coupling argument. The key ideas for the coupling argument are demonstrated in Section 4.2.4.

4.2.1. Reduction to centered random matrices when starting in equilibrium

A sample path $(Z_t)_{t=0}^\ell$ of a Markov chain Z on the state space $\mathcal{V}$ is said to have initial distribution $\iota : \mathcal{V} \to [0, 1]$ if $\mathbb{P}(Z_0 = v) = \iota(v)$ for all $v \in \mathcal{V}$. Assume that Z is irreducible and acyclic so that it has an equilibrium distribution Π_Z. Then, Z is said to *start in equilibrium* if it has initial distribution Π_Z.

Let $\hat{D}_X$ denote the $n \times n$ diagonal matrix whose ith diagonal value is the sum of the values on the ith row of $\hat{N}_X$:

$$\hat{D}_{X,ii} := \sum_{j=1}^n \hat{N}_{X,ij}. \tag{22}$$

Observe that we can write the definition of the empirical transition matrix in (3) as $\hat{P}_X = \hat{D}_X^{-1}\hat{N}_X$.

The following lemma, whose proof is provided in Section 6.3.1 based on perturbative arguments, allows us to make the following two reductions. First, we may recenter $\hat{N}_X$ and pretend as if $\hat{D}_X$ is a deterministic matrix. Second, we may assume that X starts in equilibrium.

Lemma 4.4. *Let $X := (X_t)_{t=0}^\ell$ and $Y := (Y_t)_{t=0}^\ell$ be sample paths from the block Markov chain where X starts in equilibrium and Y has an arbitrary initial distribution. Denote $M_X := \hat{N}_X - \mathbb{E}[\hat{N}_X]$ and $Q_X := \mathrm{diag}(((\ell + 1)\Pi_X)^{-1})M_X$.*

(i) *Assume that $\nu_{M_X/\sqrt{n}}$ converges weakly in probability to a probability measure ν. Then, $\nu_{\hat{N}_Y/\sqrt{n}}$ converges weakly in probability to ν.*
(ii) *Assume that $\nu_{\sqrt{n}Q_X}$ converges weakly in probability to a probability measure ν. Then, $\nu_{\sqrt{n}\hat{P}_Y}$ converges weakly in probability to ν.*

The strategy is now to apply Corollary 4.3 with $A_n = \sqrt{2}H(M_X)$ or $A_n = \sqrt{2}H(nQ_X)$ where X is a block Markov chain which starts in equilibrium.

4.2.2. Determination of the variance profile

The limiting variance profile of $\hat{N}_X$ can be established by a direct calculation which shows that the covariance between the different terms of $\hat{N}_{X,e} = \sum_{t=1}^{\ell} \mathbb{1}_{E_{X,t}=e}$ is asymptotically negligible. We instead take a different route: one that yields the stronger claim that $\hat{N}_{X,e}$ satisfies a Poisson limit theorem and is conceptually more satisfying.

Theorem 4.5. *Assume that X starts in equilibrium. Fix some $k_1, k_2 \in \{1, \ldots, K\}$ with $p_{k_1,k_2} > 0$ and let $(e_n)_{n=1}^{\infty}$ be a sequence of directed edges with $e_n \in \mathcal{V}_{k_1} \times \mathcal{V}_{k_2}$ for all n. Then $\hat{N}_{X,e_n}$ converges in distribution to a Poisson distribution with rate $\lambda \pi(k_1) \alpha_{k_1}^{-1} \alpha_{k_2}^{-1} p_{k_1,k_2}$.*

A proof is provided in Section 6.3.2 where one can also find a nonasymptotic Theorem 6.11 which gives a precise upper bound on the total variation distance of $\hat{N}_{X,e_n}$ to a Poisson distribution. The proof relies on a reduction to a general Poisson approximation theorem for sums of dependent random variables from [6].

Let us remark that the proof of Theorem 4.5 may also be used to derive a Poisson limit theorem in different scaling regimes than the running assumption $\ell = \Theta(n^2)$ and $\#\mathcal{V}_k = \Theta(n)$. More precisely, a Poisson limit theorem holds whenever ℓ and $\#\mathcal{V}_{k_1} \times \#\mathcal{V}_{k_2}$ tend to infinity in such a fashion that $(\#\mathcal{V}_{k_1} \#\mathcal{V}_{k_2})^{-1} \ell$ converges to a nonzero constant. For example, $\hat{N}_{X,e_n}$ also satisfies a Poisson limit theorem in a block Markov chain with two clusters of size $\#\mathcal{V}_1 = \Theta(1)$ and $\#\mathcal{V}_2 = \ell = \omega(1)$ respectively.

The variance profile of $\hat{N}_X$ now follows by a tightness argument which is provided in Section 6.3.3.

Corollary 4.6. *Let e_n be as in Theorem 4.5 and assume that X starts in equilibrium. Then, as n tends to infinity, it holds that $\mathrm{Var}[\hat{N}_{X,e_n}] = \lambda \pi(k_1) \alpha_{k_1}^{-1} \alpha_{k_2}^{-1} p_{k_1,k_2} + o_{k_1,k_2}(1)$.*

Graphon limits for the variance profiles of $\sqrt{2}H(M_X)$ and $\sqrt{2}H(nQ_X)$ are immediate from Corollary 4.6. To be precise, by (10) and (12) the variance profile of $\sqrt{2}H(M_X)$ converges to the graphon

$$
W_M(x, y) = \begin{cases} 2\lambda\pi(i)\alpha_i^{-1}\alpha_j^{-1}p_{i,j} & \text{if } (2x, 2y-1) \in [c_i, c_{i+1}) \times [c_j, c_{j+1}), \\ 2\lambda\pi(j)\alpha_i^{-1}\alpha_j^{-1}p_{j,i} & \text{if } (2x-1, 2y) \in [c_i, c_{i+1}) \times [c_j, c_{j+1}), \\ 0 & \text{otherwise} \end{cases}
\tag{23}
$$

with respect to the cut metric (9). Here, $c_i := \sum_{k=1}^{i-1} \alpha_k$ and $i, j \in \{1, \ldots, K\}$. Similarly, note that $\Pi_X(v) = \pi(\sigma(v))/(n\alpha_{\sigma(v)}) + o(1)$ so that the variance profile of $\sqrt{2}H(nQ_X)$ converges to the graphon W_Q specified by

$$
W_Q(x, y) = \begin{cases} 2\lambda^{-1}\pi(i)^{-1}\alpha_i\alpha_j^{-1}p_{i,j} & \text{if } (2x, 2y-1) \in [c_i, c_{i+1}) \times [c_j, c_{j+1}), \\ 2\lambda^{-1}\pi(j)^{-1}\alpha_i^{-1}\alpha_j p_{j,i} & \text{if } (2x-1, 2y) \in [c_i, c_{i+1}) \times [c_j, c_{j+1}), \\ 0 & \text{otherwise.} \end{cases}
\tag{24}
$$

4.2.3. Approximately uncorrelated

It remains to show that $H(M_X)$ and $H(nQ_X)$ are approximately uncorrelated with variance profiles. In fact, since nQ_X is derived from M_X by rescaling with $n \, \mathrm{diag}((\ell+1)\Pi_X)^{-1}$, which is a deterministic diagonal matrix with entries of size $\Theta(1)$, it is sufficient to establish that $H(M_X)$ is approximately uncorrelated with variance profile.

Proposition 4.7. *Assume that X starts in equilibrium. Then the sequence of self-adjoint random matrices $H(M_X)$ is approximately uncorrelated with variance profile.*

Recall that Definition 4.1 of approximately uncorrelated random matrices with a variance profile consists of two properties, namely (14) and (15). The proof for Proposition 4.7 given in Section 6.3.4 thus comes down to a verification of these two properties: (14) is verified in Proposition 6.15 and (15) is verified in Proposition 6.14.

The proof of Theorems 1.1 and 1.2 is then complete. Indeed, by using the Hermitian dilation in (12) and the preliminary reduction from Lemma 4.4, it is sufficient to establish limiting laws for the eigenvalues of $\sqrt{2}H(M_X)/\sqrt{2n}$ and $\sqrt{2}H(nQ_X)/\sqrt{2n}$ when X starts in equilibrium. This case follows from Corollary 4.3 with the limiting variance profiles in (23) and (24).

4.2.4. Demonstration of the coupling argument

Proposition 4.7 is the most important ingredient for our results. Let us provide an example for the method of proof by establishing a special case of (14): the covariance between two entries decays at an appropriate rate.

Proposition 4.8. *Assume that X starts in equilibrium. Then,*

$$\max_{e_1 \neq e_2} |\mathbb{E}[M_{X,e_1} M_{X,e_2}]| = O(n^{-1}) \tag{25}$$

where the maximum runs over all pairs of distinct edges $e_1, e_2 \in \vec{E}_n$.

Proof. The proof is split into parts. The main ideas are contained in Part 2 and Part 3. In Part 2 we observe that it is sufficient to understand how much the expectation of $\hat{N}_{X,e_2}$ changes when it is conditioned on a traversal of e_1 at some predetermined time. This effect of conditioning on a traversal is then understood by a coupling argument in Part 3.

Part 1: Preliminary reduction to $K \geq 5$

We claim that there is no loss in generality in assuming that $K \geq 5$. The idea is to split a cluster into pieces.

Since $\#\mathcal{V}_K = \alpha_K n + o(n)$ it may be assumed that $\#\mathcal{V}_K \geq 5$. Define a new partition of the state space $\mathcal{V}$ into nonempty sets $(\mathcal{V}'_i)_{i=1}^{K+4}$ by $\mathcal{V}'_i := \mathcal{V}_i$ for $i < K$ and by taking $(\mathcal{V}'_i)_{i=K}^{K+4}$ to be a partition of $\mathcal{V}_K$ into nonempty sets. It can here be ensured that we remain in the asymptotic regime where the clusters have size $\Theta(n)$. Indeed, this for instance follows if the subdivision of $\mathcal{V}_K$ is taken to be into clusters of roughly equal size so that the ratio $\#\mathcal{V}'_i/\#\mathcal{V}_K$ tends to $1/5$ for every $i \in \{K, \ldots, K+4\}$. Further define a $(K+4) \times (K+4)$ stochastic matrix p' by $p'_{i,j} := (\#\mathcal{V}'_j/\#\mathcal{V}_{\min\{K,j\}})p_{\min\{K,i\},\min\{K,j\}}$ for all $i, j = 1, \ldots, K+4$.

The reduction to $K \geq 5$ now follows by observing that the clusters $(\mathcal{V}'_i)_{i=1}^{K+4}$ and cluster transition matrix p' define exactly the same block Markov chain as we started with. Indeed, if $P'_{i,j}$ denotes the transition probabilities of the 'new' block Markov chain then for any $i \in \mathcal{V}'_x$ and $j \in \mathcal{V}'_y$ it holds that

$$P'_{i,j} = \frac{1}{\#\mathcal{V}'_y} p'_{x,y} = \frac{1}{\#\mathcal{V}'_y} \frac{\#\mathcal{V}'_y}{\#\mathcal{V}_{\min\{K,y\}}} p_{\min\{K,x\},\min\{K,y\}} = P_{i,j}. \tag{26}$$
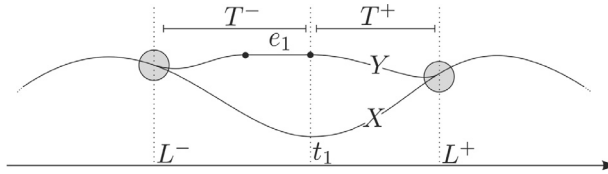
Fig. 3. Visualization of the construction of the chain Y in the proof of Proposition 4.8 by gluing an edge e_1 onto X at time t_1. The clusters of the block Markov chain allow us to ensure that $\mathbb{E}[L^+ - L^-]$ is $O(1)$ which causes the expected difference $\mathbb{E}[\hat{N}_{Y,e_2} - \hat{N}_{X,e_2}]$ to be small because a traversal of e_2 in this small period of time is unlikely.

Part 2: Rewriting $\mathbb{E}[M_{X,e_1} M_{X,e_2}]$

Pick two distinct edges $e_1, e_2 \in \vec{E}_n$. Recall from Section 3.1 that $E_X = (E_{X,t})_{t=1}^{\ell}$ denotes the induced Markov chain of edges $E_{X,t} = (X_{t-1}, X_t)$. It may be assumed that e_1 is such that $\mathbb{P}(E_{X,1} = e_1) \neq 0$, otherwise $M_{X,e_1} = 0$ and there is nothing to prove.

Recall that $M_X = \hat{N}_X - \mathbb{E}[\hat{N}_X]$ and write $\hat{N}_{X,e_1} = \sum_{t_1=1}^{\ell} \mathbb{1}_{E_{X,t_1}=e_1}$ to find that

$$\mathbb{E}[M_{X,e_1} M_{X,e_2}] = \mathbb{E}[\hat{N}_{X,e_1} M_{X,e_2}] - \mathbb{E}[\hat{N}_{X,e_1}]\mathbb{E}[M_{X,e_2}] \tag{27}$$

$$= \sum_{t_1=1}^{\ell} \mathbb{P}(E_{X,t_1} = e_1)\left(\mathbb{E}[\hat{N}_{X,e_2} \mid E_{X,t_1} = e_1] - \mathbb{E}[\hat{N}_{X,e_2}]\right). \tag{28}$$

Note that all edges whose starting point and ending point have the same clusters as the starting point and ending point of e_1 respectively are equally likely to be traversed at time t_1. There are at least $\alpha_{min}^2 n^2$ such edges. Hence, $\mathbb{P}(E_{X,t_1} = e_1) \leq \alpha_{min}^{-2} n^{-2}$. Considering that there are $\ell = \Theta(n^2)$ terms on the right-hand side of (28), it remains to be shown that $\mathbb{E}[\hat{N}_{X,e_2} \mid E_{X,t_1} = e_1] - \mathbb{E}[\hat{N}_{X,e_2}] = O(n^{-1})$ uniformly in t_1.

Part 3: Construction of coupled chains (X, Y)

Recall that we ensured that $K \geq 5$. In particular there exists some $k \in \{1, \ldots, K\}$ such that $\mathcal{V}_k$ does not contain any endpoint of e_1 and e_2. To study the difference $\mathbb{E}[\hat{N}_{X,e_2} \mid E_{X,t_1} = e_1] - \mathbb{E}[\hat{N}_{X,e_2}]$ we construct a pair of chains (X, Y); see Fig. 3 for a visualization.

(i) Sample an infinitely long path $\widetilde{X} := (\widetilde{X}_t)_{t=-\infty}^{\infty}$ from the block Markov chain and independently sample an infinitely long path $\widetilde{Y} := (\widetilde{Y}_t)_{t=-\infty}^{\infty}$ from the block Markov chain conditioned on $E_{\widetilde{Y},t_1} = e_1$. Note that it is possible to sample at negative times by means of a time reversal of the Markov chain. Such time reversal exists by the assumption that the Markov chain associated with p is irreducible and acyclic.

(ii) Define

$$T^- := t_1 - \sup\{t \in \mathbb{Z}_{<t_1} : \widetilde{X}_t \in \mathcal{V}_k, \widetilde{Y}_t \in \mathcal{V}_k\}, \tag{29}$$

$$T^+ := \inf\{t \in \mathbb{Z}_{>t_1} : \widetilde{X}_t \in \mathcal{V}_k, \widetilde{Y}_t \in \mathcal{V}_k\} - t_1 \tag{30}$$

and note that T^- and T^+ are finite with probability one due to the assumption that the Markov chain associated with p is irreducible and acyclic. Let $L^- := \max\{0, t_1 - T^-\}$ and $L^+ := \min\{\ell, t_1 + T^+\}$.

(iii) Let $X := (\widetilde{X}_t)_{t=0}^{\ell}$. Define $Y_t := (Y_t)_{t=0}^{\ell}$ by $Y_t = \widetilde{Y}_t$ for $t \in \{L^-, \ldots, L^+\}$ and $Y_t = \widetilde{X}_t$ otherwise.

By construction, X is a sample path from the block Markov chain whereas Y is a sample path of the block Markov chain conditioned on the event $E_{Y,t_1} = e_1$. Now, by the law of total

expectation

$$\mathbb{E}[\hat{N}_{X,e_2} \mid E_{X,t_1} = e_1] - \mathbb{E}[\hat{N}_{X,e_2}] = \mathbb{E}[\hat{N}_{Y,e_2} - \hat{N}_{X,e_2}] \tag{31}$$

$$= \sum_{\ell^-, \ell^+=1}^{\ell} \mathbb{P}(L^- = \ell^-, L^+ = \ell^+)\mathbb{E}[\hat{N}_{Y,e_2} - \hat{N}_{X,e_2} \mid L^- = \ell^-, L^+ = \ell^+]. \tag{32}$$

Part 4: $|\mathbb{E}[\hat{N}_{Y,e_2} - \hat{N}_{X,e_2} \mid L^- = \ell^-, L^+ = \ell^+]| \le 2\alpha_{min}^{-1} n^{-1}(\ell^+ - \ell^-)$

Let $\Delta_X := \sum_{t=L^-+1}^{L^+} \mathbb{1}_{E_{X,t}=e_2}$ denote the number of times e_2 was traversed by $(X_t)_{t=L^-}^{L^+}$. Similarly, let $\Delta_Y := \sum_{t=L^-+1}^{L^+} \mathbb{1}_{E_{Y,t}=e_2}$. Since $\mathcal{V}_k$ does not contain any endpoint of e_2, it holds that $\hat{N}_{Y,e_2} - \hat{N}_{X,e_2} = \Delta_Y - \Delta_X$. Therefore

$$|\mathbb{E}[\hat{N}_{Y,e_2} - \hat{N}_{X,e_2} \mid L^- = \ell^-, L^+ = \ell^+]| \tag{33}$$
$$\le \mathbb{E}[\Delta_X \mid L^- = \ell^-, L^+ = \ell^+] + \mathbb{E}[\Delta_Y \mid L^- = \ell^-, L^+ = \ell^+].$$

We will establish a bound on the conditional expectation of Δ_Y by using its definition in terms of $\mathbb{1}_{E_{Y,t}=e_2}$. To this end we claim that

$$\mathbb{P}(E_{Y,t} = e_2 \mid L^- = \ell^-, L^+ = \ell^+) \le \alpha_{min}^{-1} n^{-1} \tag{34}$$

for any $t \in \{1, \ldots, \ell\}$. In case $t = t_1$, the left-hand side of (34) is zero and there is nothing to prove. Now consider the case where $t \ne t_1$. For any edge e whose starting point is equal to the starting point of e_2 and whose ending point is in the same cluster as the ending point of e_2 it holds that $\mathbb{P}(E_{Y,t} = e \mid L^- = \ell^-, L^+ = \ell^+) = \mathbb{P}(E_{Y,t} = e_2 \mid L^- = \ell^-, L^+ = \ell^+)$. This implies (34) whenever $t > t_1$ since there are at least $\alpha_{min} n$ such edges e. The case $t < t_1$ may be deduced similarly by reversing the roles of the ending point and the starting point of e.

Combine (34) with the fact that $\Delta_Y = \sum_{t=L^-+1}^{L^+} \mathbb{1}_{E_{Y,t}=e_2}$ to conclude that

$$\mathbb{E}[\Delta_Y \mid L^- = \ell^-, L^+ = \ell^+] = \sum_{t=\ell^-+1}^{\ell^+} \mathbb{P}(E_{Y,t} = e_2 \mid L^- = \ell^-, L^+ = \ell^+) \tag{35}$$

$$\le \alpha_{min}^{-1} n^{-1}(\ell^+ - \ell^-). \tag{36}$$

The same conclusion applies to Δ_X. Combine finally with (32) and (33) to deduce that

$$|\mathbb{E}[\hat{N}_{X,e_2} \mid E_{X,t_1} = e_1] - \mathbb{E}[\hat{N}_{X,e_2}]| \le 2\alpha_{min}^{-1} n^{-1}\mathbb{E}[L^+ - L^-]. \tag{37}$$

Part 5: $\mathbb{E}[L^+ - L^-] = O(1)$

Note that $\mathbb{E}[L^+ - L^-] \le \mathbb{E}[T^+] + \mathbb{E}[T^-]$. Here, $\mathbb{E}[T^+]$ and $\mathbb{E}[T^-]$ are finite constants which do not depend on n. Indeed, consider the product Markov chain $\Sigma_{(X,Y)}^+ := (\sigma(X_{t_1+t}), \sigma(Y_{t_1+t}))_{t=0}^{\infty}$ on the space $\{1, \ldots, K\} \times \{1, \ldots, K\}$. Then T^+ is the first strictly positive time $\Sigma_{(X,Y)}^+$ is in (k, k). Recall that the transition dynamics p for $\sigma(X_t)$ and $\sigma(Y_t)$ are assumed to be acyclic and irreducible. It follows that $\mathbb{P}(T^+ > t)$ shows exponential decay in t. Because there are only K^2 possibilities for the initial state $\Sigma_{(X,Y),0}^+$ it follows that $\mathbb{E}[T^+] \le B^+$ for some constant $B^+ \in \mathbb{R}_{>0}$ which does not depend on e_1, e_2 or n. A similar argument shows that $\mathbb{E}[T^-] \le B^-$ for some $B^- \in \mathbb{R}_{>0}$.

From (28) and (37) it now follows that

$$\mathbb{E}[M_{X,e_1} M_{X,e_2}] \le 2\ell(B^+ + B^-)\alpha_{min}^{-3} n^{-3}. \tag{38}$$

Observe that the right-hand side of (38) is independent of e_1, e_2. Since $\ell = \Theta(n^2)$ this concludes the proof. □

5. Numerical experiment on Manhattan taxi trips

We will now demonstrate that Theorems 1.1 and 1.2 can give nontrivial predictions for the singular value distributions on an actual dataset. Specifically, we will analyze the first six months of 2016 in the New York City yellow cab dataset [57]. Each datapoint contains the pick-up and drop-off location of one trip. Here, pick-up locations are typically close to drop-off locations so that the dataset may be modeled as a fragmented sample path of a Markov chain.

Spectral clusterings using the Markovian structure of this dataset have previously been analyzed in [77]. Our preprocessing is similar to what was done in [77], and is as follows. The map is subdivided into a fine grid and we trim all states which have been visited fewer than 200 times. We further remove all self-transitions. This results in a state space of size $n = 4486$ with $\ell = 55 \times 10^6$ transitions.

Note that $\ell/n^2 \approx 2.7$. This empirical observation allows us to make a relevant remark concerning our theoretical assumption $\ell = \Theta(n^2)$. Some of our readers may namely be familiar with the literature on random graphs. It is an empirical observation that most real-world graphs are sparse. This sparsity is correspondingly a key difficulty which one should interact with in the setting of random graphs. Sparsity is not irrelevant in the setting of sequential data but it has a different meaning; it relates to the amount of time that the process was observed. As opposed to random graphs it is not unusual to encounter dense sequential data in the real world. The key novel difficulty is rather that sequential data has dependence. This is precisely the difficulty which our proofs interact with; recall the coupling argument in the proof of Proposition 4.8.

A clustering $(\hat{\mathcal{V}}_k)_{k=1}^K$ is found by applying both steps from the algorithm in [66] with $K = 4$ clusters; the result is displayed on the left-hand side of Fig. 4. Having obtained these clusters we may estimate the parameters of the block Markov chain as

$$\hat{\lambda} = \frac{\ell}{n^2}, \quad \hat{\alpha}_k = \frac{\#\hat{\mathcal{V}}_k}{n}, \quad \hat{\pi}_k = \frac{\sum_{i \in \mathcal{V}} \sum_{j \in \hat{v}_k} \hat{N}_{X,ij}}{\sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} \hat{N}_{X,ij}}, \quad \hat{p}_{k_1,k_2} = \frac{\sum_{i \in \hat{v}_{k_1}} \sum_{j \in \hat{v}_{k_2}} \hat{N}_{X,ij}}{\sum_{i \in \hat{v}_{k_1}} \sum_{j \in \mathcal{V}} \hat{N}_{X,ij}}.$$

These parameters may be substituted in Theorems 1.1 and 1.2 to yield predictions for the singular value distributions of $\hat{N}_X$ and $\hat{P}_X$. The theoretical predictions and the empirical observations are displayed in Fig. 4. For comparison, we have also displayed the quarter circle law which is the universal law for the singular values of a random matrix with independent entries and identical variance. In other words, the quarter circle law is the prediction corresponding to $K = 1$.

Taking into account the fact that we used just $K = 4$ clusters, we conclude that the predictions match the shape of the singular value distributions fairly well. Observe that the quarter circle law does not even match the general shape of the distributions: the quarter circle law has a concave density whereas the observed empirical distributions have convex densities.

6. Proofs

6.1. Proof of Theorem 4.2

We work under the assumptions of Theorem 4.2. This is to say that $(A_n)_{n=1}^\infty$ is a family of symmetric $n \times n$ random matrices which are approximately uncorrelated with variance profile.
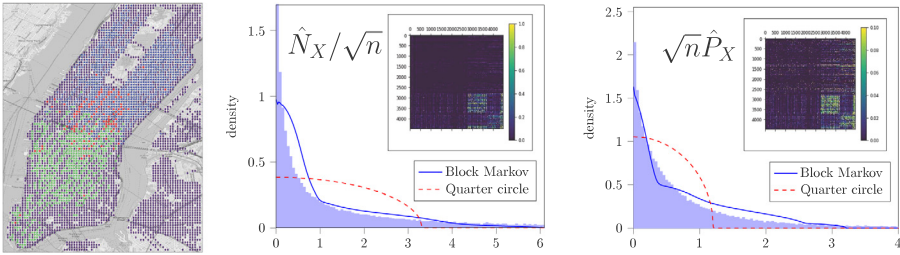
Fig. 4. On the left: Manhattan and the four clusters in purple, blue, green and red. In the middle: $\hat{N}_X/\sqrt{n}$ and a frequency-based histogram of the singular values compared to our theoretical predictions and the quarter-circle law with density $(\pi\hat{\lambda})^{-1}(4\hat{\lambda} - x^2)^{1/2}\mathbb{1}_{x\in[0,2\sqrt{\hat{\lambda}}]}$. On the right: $\sqrt{n}\hat{P}_X$ and its singular values compared to our theoretical predictions and the quarter-circle law with density $\pi^{-1}\hat{\lambda}(4/\hat{\lambda} - x^2)^{1/2}\mathbb{1}_{x\in[0,2/\sqrt{\hat{\lambda}}]}$. The systems of equations in Theorems 1.1 and 1.2 were solved using the algorithm in [39, Proposition 4.1] after which we used the Stieltjes inversion formula (4) to recover the measures.

Recall that for any fixed ordered tree $T \in \mathcal{T}_k$ it is assumed that $t(T, S_n)$ has a limit as $n \to \infty$ where $S_n = (\mathrm{Var}(A_{n,ij}))_{i,j=1}^n$ denotes the variance profile of A_n.

The proof of Theorem 4.2 comes down to a modification of the proof in [40] to include the variance profile. Let us start by including some background on the moment method. These results are well-known and included for the reader's convenience.

The following lemma is implicit in the proof of the Wigner semicircle law in [5, Section 2.1]. See particularly the remarks following Lemma 2.1.7 in [5, Section 2.1.2] and the application of Chebyshev's inequality in the first sentence of [5, Section 2.1.4].

Lemma 6.1. *Let $(\mu_n)_{n=1}^\infty$ be a sequence of random probability measures on $\mathbb{R}$ and let μ be a deterministic and compactly supported probability measure on $\mathbb{R}$. If, for every $k \in \mathbb{Z}_{\geq 0}$,*

$$\mathbb{E}[\mathfrak{m}_k(\mu_n)] = \mathfrak{m}_k(\mu) + o_k(1); \quad \mathrm{Var}[\mathfrak{m}_k(\mu_n)] = o_k(1), \tag{39}$$

then μ_n converges weakly in probability to μ.

The following result is moreover a direct consequence of the Stone–Weierstrass theorem and the Riesz representation theorem [65, Theorem 2.14].

Lemma 6.2. *For any compactly supported probability measure μ it holds that the sequence of moments $(\mathfrak{m}_k(\mu))_{k=0}^\infty$ satisfies the following properties:*

  (i) *It holds that $\mathfrak{m}_0(\mu) = 1$.*
  (ii) *For any $k \in \mathbb{Z}_{\geq 0}$ the Hankel matrix $(\mathfrak{m}_{i+j}(\mu))_{i,j=0}^k$ is positive semi-definite.*
  (iii) *There exists a constant $c \in \mathbb{R}_{>0}$ such that $|\mathfrak{m}_k(\mu)| \leq c^k$ for all $k \in \mathbb{Z}_{\geq 0}$.*

Moreover, for any sequence of real number $(m_k)_{k=0}^\infty$ satisfying these properties there exists a unique probability measure μ with $(\mathfrak{m}_k(\mu))_{k=0}^\infty = (m_k)_{k=0}^\infty$ and this measure μ is compactly supported.

Lemma 6.3. *There exists a unique probability measure μ whose moments are given by*

$$\mathfrak{m}_{2m}(\mu) = \sum_{T \in \mathcal{T}_m} \lim_{n\to\infty} t(T, S_n); \quad \mathfrak{m}_{2m+1}(\mu) = 0, \tag{40}$$

for every $m \in \mathbb{Z}_{\geq 0}$. Moreover, μ is compactly supported.

Proof. The existence of such a probability measure μ is known [79, Theorem 3.2]. However, it is not explicitly stated in [79] that μ is compactly supported so, for the sake of completeness, let us provide an argument based on Lemma 6.2.

The condition that $\mathfrak{m}_0(\mu) = 1$ is satisfied since μ is a probability measure. Further, the positive semi-definiteness of the Hankel matrix $(\mathfrak{m}_{i+j}(\mu))_{i,j=0}^k$ for every $k \in \mathbb{Z}_{\geq 0}$ is equivalent to the trivial statement that $\int p(x)\overline{p(x)}d\mu(x) \geq 0$ for every polynomial $p(x) \in \mathbb{C}[x]$. It remains to show that the moments of μ are exponentially bounded.

It follows from (14) in the definition of an approximately uncorrelated random matrix with a variance profile that $\max_{i,j=1,\ldots,n} S_{n,ij} \leq c_1$. Therefore, using the definition of homomorphism densities in (16) it holds that for every $m \in \mathbb{Z}_{\geq 0}$ and any ordered tree $T \in \mathcal{T}_m$,

$$t(T, S_n) \leq c_1^m. \tag{41}$$

It holds that $\#\mathcal{T}_m = C_m$ where C_m is the mth Catalan number. It is known that there exists some $c_2 \in \mathbb{R}_{>0}$ such that $C_m \leq c_2^m$ for all $m \in \mathbb{Z}_{\geq 0}$. Hence,

$$\sum_{T \in \mathcal{T}_m} t(T, S_n) \leq c_2^m c_1^m \leq c_3^{2m} \tag{42}$$

for some constant $c_3 \in \mathbb{R}_{>0}$ and all $m \geq 1$. (42) provides an exponential bound on the rate of growth of the even moments $\mathfrak{m}_{2m}(\mu)$. Further, recall that we already know that $\mathfrak{m}_0(\mu) = 1$ and $\mathfrak{m}_{2m+1} = 0$. It follows that $\mathfrak{m}_k(\mu) \leq c_3^k$ for all $k \in \mathbb{Z}_{\geq 0}$. Conclude by Lemma 6.2 that μ is compactly supported and unique. $\square$

We adapt the same notation as in the sketch of Theorem 4.2. This is to say that given integers $i := (i_1, \ldots, i_k, i_1)$ with $i_j \in \{1, \ldots, n\}$ for every $j = 1, \ldots, k$ we denote the induced undirected graph with vertex set $V(i) := \{i_1, \ldots, i_k\}$ and edge set $E(i) := \{\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_k, i_1\}\}$ by $G_i = (V(i), E(i))$. Viewing i as a cycle on G_i we let $r_1(i)$ be the number of edges which are traversed exactly once and $r_2(i)$ be the number of edges which are traversed exactly twice. Finally, we define $P(i) := \mathbb{E}[A_{n,i_1 i_2} \cdots A_{n,i_k i_1}]$.

The following combinatorial result will be essential.

Lemma 6.4 ([40, Corollary 10]). *With i as above it holds that $\#V(i) \leq (k + r_1(i))/2 + 1$ with strict inequality whenever $r_1(i) > 0$.*

Lemma 6.5. *For any $m \in \mathbb{Z}_{\geq 0}$ it holds that*

$$\mathbb{E}[\mathfrak{m}_{2m+1}(\mu_{A_n/\sqrt{n}})] = o_m(1), \tag{43}$$

$$\mathbb{E}[\mathfrak{m}_{2m}(\mu_{A_n/\sqrt{n}})] = \sum_{T \in \mathcal{T}_m} t(T, S_n) + o_m(1). \tag{44}$$

Proof. Let $k \in \mathbb{Z}_{\geq 0}$ be a positive integer and recall the expansion for $\mathbb{E}[\mathfrak{m}_k(\mu_{A_n/\sqrt{n}})]$ in (19). Let $i := (i_1, \ldots, i_k, i_1)$ be a sequence of integers as occurs on the right-hand side of (19). By property (14) in the definition of an approximately uncorrelated random matrix it holds that $P(i) = O_k(n^{-r/2})$ whenever $r_1(i) = r$. By the part concerning the strict inequality in Lemma 6.4 we have that for any $r \in \mathbb{Z}_{>0}$ there are $o_k(n^{(k+r)/2+1})$ terms on the right-hand side of (19) with $r_1(i) = r$. This means that only the terms with $r_1(i) = 0$ survive the normalization by $n^{-1-k/2}$:

$$\mathbb{E}[\mathfrak{m}_k(\mu_{A_n/\sqrt{n}})] = n^{-1-k/2} \sum_{i:r_1(i)=0} P(i) + o_k(1). \tag{45}$$
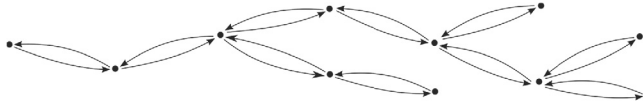
Fig. 5. Visualization of a sequence of integers i for which $P(i)$ has an asymptotically relevant contribution to (45). The corresponding undirected graph G_i is the tree found by identifying the doubled edges.

Note that G_i is connected. Hence, for any i with $r_1(i) = 0$ it has to hold that $\#V(i) \leq k/2 + 1$ with equality if and only if k is even, G_i is a tree and $r_2(i) = k/2$. In particular, for k odd there are $o_k(n^{1+k/2})$ terms on the right-hand side of (45). Hence, for any $m \in \mathbb{Z}_{\geq 0}$ taking $k = 2m + 1$ yields that

$$\mathbb{E}[\mathfrak{m}_{2m+1}(\mu_{A_n/\sqrt{n}})] = o_m(1). \tag{46}$$

Now let $k = 2m$ be even. As above, the contribution of the terms for which G_i is not a tree or $r_2(i) \neq m$ is asymptotically negligible. Consider some sequence of indices $i = (i_1, \ldots, i_k, i_1)$ for which G_i is a tree and $r_2(i) = m$. An example of such a sequence i is depicted in Fig. 5. Equip G_i with the unique order such that i is the path traversed by depth-first search.

Now,

$$\mathbb{E}[\mathfrak{m}_{2m}(\mu_{A_n/\sqrt{n}})] = n^{-1-(2m)/2} \sum_{T \in \mathcal{T}_m} \sum_{i : G_i \cong T} P(i) + o_m(1) \tag{47}$$

where the isomorphism condition is to be considered in the space of ordered trees.

The ordering on any $T \in \mathcal{T}_m$ induces a canonical numbering of the vertex set corresponding to the order of visits in depth-first search. This is to say that it can be assumed that $V(T) = \{1, \ldots, m+1\}$. Denote $E(T)$ for the set of edges. Note that the collection of sequences of indices i with $G_i \cong T$ is in bijection with the collection of injective labelings $l^i : V(T) \to \{1, \ldots, n\}$. The assumption that A_n is approximately uncorrelated with variance profile S_n then implies that $P(i) = \prod_{\{v,w\} \in E(T)} S_{n,l_v^i l_w^i} + o_m(1)$. Note that there are $O_m(n^{m+1})$ injective labelings of $V(T)$ and only $O_m(n^m)$ labelings which are not injective. Hence, the sum over all i with $G_i \cong T$ is asymptotically equivalent to a sum over all (not necessarily injective) labelings $l : V(T) \to \{1, \ldots, n\}$:

$$n^{-1-(2m)/2} \sum_{i : G_i \cong T} P(i) = n^{-1-(2m)/2} \sum_{l_1, \ldots, l_{m+1} = 1}^{n} \prod_{\{v,w\} \in E(T)} S_{n,l_v l_w} + o_m(1) \tag{48}$$

$$= t(T, S_n) + o_m(1). \tag{49}$$

Note that (16) was used in the second step. By (47) it now follows that for every $m \in \mathbb{Z}_{\geq 0}$

$$\mathbb{E}[\mathfrak{m}_{2m}(\mu_{A_n/\sqrt{n}})] = \sum_{T \in \mathcal{T}_m} t(T, S_n) + o_m(1). \tag{50}$$

Combine (46) and (50) to conclude the proof. $\quad\square$

Lemma 6.6. *For any $k \in \mathbb{Z}_{\geq 0}$ it holds that* $\operatorname{Var}[\mathfrak{m}_k(\mu_{A_n/\sqrt{n}})] = o_k(1)$.
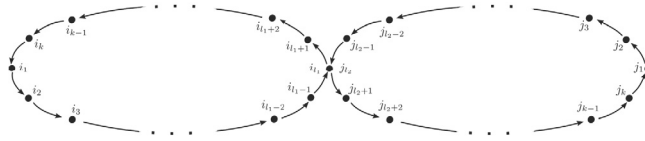
Fig. 6. Visualization of a pair of sequences of integers i, j with $V(i) \cap V(j) \neq \emptyset$ as occur in the proof of Lemma 6.6. While not depicted it is possible that i and j are also equal at other vertices than $i_{\ell_1} = j_{\ell_2}$.

Proof. Recall that

$$\mathrm{Var}[\mathfrak{m}_k(\mu_{A_n/\sqrt{n}})] = \mathbb{E}[\mathfrak{m}_k(\mu_{A_n/\sqrt{n}})^2] - \mathbb{E}[\mathfrak{m}_k(\mu_{A_n/\sqrt{n}})]^2. \tag{51}$$

The limit of $\mathbb{E}[\mathfrak{m}_k(\mu_{A_n/\sqrt{n}})]^2$ is known by Lemma 6.5. Hence, it suffices to show that $\mathbb{E}[\mathfrak{m}_k(\mu_{A_n/\sqrt{n}})^2]$ converges to the same limit. For any $k \in \mathbb{Z}_{\geq 0}$,

$$\mathbb{E}[\mathfrak{m}_k(\mu_{A_n/\sqrt{n}})^2] = n^{-2-k}\mathbb{E}[(\mathrm{Tr}\, A_n^k)^2] \tag{52}$$

$$= n^{-2-k} \sum_{i_1,\ldots,i_k=1}^{n} \sum_{j_1,\ldots,j_k=1}^{n} \mathbb{E}[A_{n,i_1 i_2} A_{n,i_2 i_3} \cdots A_{n,i_k i_1} A_{n,j_1 j_2} A_{n,j_2 j_3} \cdots A_{n,j_k j_1}]. \tag{53}$$

Let $i := (i_1, \ldots, i_k, i_1)$ and $j := (j_1, \ldots, j_k, j_1)$ be sequences of integers which occur in (53). Define $G_{i,j} := (V(i) \cup V(j), E(i) \cup E(j))$. Let $P(i,j) := \mathbb{E}[A_{n,i_1 i_2} \cdots A_{n,j_k j_1}]$ be the term occurring on the right-hand side of (53).

First, consider those sequences of indices i, j for which $V(i) \cap V(j) \neq \emptyset$. Then i, j can be merged: let $l_1 \in \{1, \ldots, k\}$ be the minimal index such that $i_{l_1} \in \{j_1, \ldots, j_r\}$ and let $l_2 \in \{1, \ldots, k\}$ be the minimal index such that $j_{l_2} = i_{l_1}$ and set

$$\ell_{i,j} := (i_1, i_2, \ldots, i_{l_1}, j_{l_2+1}, \ldots, j_k, j_1, j_2, \ldots, j_{l_2}, i_{l_1+1}, \ldots, i_k, i_1). \tag{54}$$

A resulting sequence is schematically depicted in Fig. 6.

By definition of approximately uncorrelated it holds that $P(i,j) = O_k(n^{-r/2})$ whenever $r_1(\ell_{i,j}) = r$. Further, it follows from Lemma 6.4 that $\#(V(i) \cup V(j)) \leq (2k + r_1(\ell_{i,j}))/2 + 1$. Hence, for any $r \in \mathbb{Z}_{\geq 0}$ there are $O_k((2k+r)/2+1)$ terms with $V(i) \cap V(j) \neq \emptyset$ and $r_1(\ell_{i,j}) = r$. This implies that the contribution is asymptotically negligible:

$$n^{-2-k} \sum_{i,j:V(i)\cap V(j)\neq\emptyset} P(i,j) = O_k(n^{-1}). \tag{55}$$

The argument to deal with the terms with $V(i) \cap V(j) = \emptyset$ is identical to the argument used in the proof of Lemma 6.5. In particular, it can be established that

$$n^{-2-k} \sum_{i,j:V(i)\cap V(j)=\emptyset} P(i,j) = \begin{cases} o_k(1) & \text{if } k \text{ is odd,} \\ \left(\sum_{T \in \mathcal{T}_{k/2}} t(T, S_n)\right)^2 + o_k(1) & \text{if } k \text{ is even,} \end{cases} \tag{56}$$

where it is used that the contribution of those pairs of trees G_i, G_j which were removed by restricting our attention to the case of $V(i) \cap V(j) = \emptyset$ is asymptotically negligible due to the fact that there are only $O_k(n^{k+1})$ values of i, j with $V(i) \cap V(j) \neq \emptyset$ and $r_2(i) = r_2(j) = k/2$.

Combine (55), (56), and the limit established in Lemma 6.5 for $\mathbb{E}[\mathfrak{m}_k(\mu_{A_n/\sqrt{n}})]^2$ to deduce that $\mathrm{Var}[\mathfrak{m}_k(\mu_{A_n/\sqrt{n}})] = o_k(1)$. $\square$

Proof of Theorem 4.2. By Lemma 6.3 there exists a compactly supported probability measure μ with the specified moments. The result is now immediate by Lemma 6.1 whose assumptions were verified in Lemmas 6.5 and 6.6. $\square$

6.2. *Proof of Corollary 4.3*

Proof. Recall that it is assumed that $\delta_{\square}(W^{S_n}, W) \to 0$. Then, by [51, Theorem 11.5], it holds that $t(F, W^{S_n})$ converges as $n \to \infty$ for every fixed tree. Consequently, by Theorem 4.2, it holds that $\mu_{A_n/\sqrt{n}}$ converges weakly in probability to some limit μ and it remains to show that the Stieltjes transform of μ is given by (21).

To this end remark that Theorem 4.2 also applies to random matrices with independent entries. Consequently, if we consider a sequence of random matrices B_n with independent entries and the same variance profile as A_n, then also $\mu_{B_n/\sqrt{n}}$ converges weakly in probability to μ. The desired shape for the Stieltjes transform now follows from Theorem 4.2 by [79, Theorem 3.4] which provides the same description for the Stieltjes transform in the case of a random matrix with independent entries. The assumption that $\max_{ij=1,\dots,n} S_{n,ij} = O(1)$ in [79] is satisfied since (14) in the definition of an approximately uncorrelated random matrix implies that $\max_{i,j=1,\dots,n} \mathbb{E}[A_{n,ij}^2] = O(1)$.

Let us remark that it may not be clear from the statement of [79, Theorem 3.4.] that the codomain of $a(z, x)$ should be $\mathbb{C}^-$. This specification of codomain is however important to guarantee uniqueness of the solution to the self-consistent equation (21); see [3, Theorem 2.1 and subsequent remarks]. The fact that the codomain should be $\mathbb{C}^-$ is implicit in the final sentence of the proof of [79, Theorem 3.4] which makes reference to [3, Theorem 2.1]. Let us provide some additional details for this final sentence of the proof in [79] so as to make the codomain explicit.

The proof of [79, Theorem 3.4] begins by specifying the function by means of a Laurent series, namely $a(z, x) := \sum_{k=0}^{\infty} \beta_{2k}(x)z^{-2k-1}$ for certain explicit coefficients $\beta_{2k}(x)$, in [79, Equation (3.4)] after which it is deduced that (20) and (21) are satisfied by the provided Laurent series. On the other hand, by using a geometric series at $z = \infty$ in [3, (2.8)], one can see that the unique solution $\widetilde{a}(z, x)$ to (21) with codomain $\mathbb{C}^-$ also admits a Laurent series, namely $\widetilde{a}(z, x) = \sum_{k=1}^{\infty} b_k(x)z^{-k}$. Let us warn here that the convention for the Stieltjes transform which was used in [3] differs by a minus sign from the convention which was used here and in [79]. One should correspondingly take $m_x(z) = -a(z, x)$ when applying [3, Theorem 2.1]. This difference by a sign also explains why the codomain in [3] is $\mathbb{C}^+$ whereas we claim that the appropriate codomain is $\mathbb{C}^-$.

Now observe that a Laurent series of the form $\sum_{k=1}^{\infty} c_k(x)z^{-k}$ satisfies (21) if and only if $c_1(x) = 1$, $c_2(x) = 0$, and

$$c_k(x) = \int_0^1 W(x, y) \sum_{j=1}^{k-2} c_{k-j-1}(x)c_j(y)\mathrm{d}y \quad \text{for all } k \geq 3. \tag{57}$$

This implies that the coefficients of the Laurent series of $a(z, x)$ and $\widetilde{a}(z, x)$ have to be equal since both functions satisfy (21). It follows that $a(z, x)$ and $\widetilde{a}(z, x)$ are equal in a neighborhood of $z = \infty$ and consequently everywhere on $\mathbb{C}^+$ by the identity theorem for analytic functions. This shows that the function $a(z, x)$ provided in [79, Equation (3.4)] indeed has codomain $\mathbb{C}^-$. $\square$

6.3. Proofs for Theorems 1.1 and 1.2

It remains to establish the claims which were announced in the proof outline in Section 4.2. This means that we have to prove the following statements: (i) The preliminary reduction to centered random matrices starting in equilibrium in Lemma 4.4. (ii) The Poisson limit Theorem 4.5 and its Corollary 4.6 concerning the variance profile of $\hat{N}_X$. (iii) The statement that $H(M_X)$ is approximately uncorrelated with variance profile from Proposition 4.7.

6.3.1. Proof of preliminary reductions in Lemma 4.4

Recall from Section 3.4 that ν_A denotes the empirical singular value distribution of a matrix A.

Lemma 6.7. *Let A_n be a sequence of random $n \times n$ matrices such that ν_{A_n} converges weakly in probability to some probability measure ν.*

(i) *Let B_n be a sequence of random $n \times n$ matrices such that $\frac{1}{n}\|B_n\|_{\mathrm{F}}^2$ converges to 0 in probability. Then $\nu_{A_n+B_n}$ converges weakly in probability to ν.*

(ii) *Let C_n be a sequence of random symmetric $n \times n$ matrices such that $\frac{1}{n}\mathrm{rank}(C_n)$ converges to 0 in probability. Then $\nu_{A_n+C_n}$ converges weakly in probability to ν.*

(iii) *Let D_n be a sequence of random diagonal $n \times n$ matrices such that $\|D_n - \mathrm{Id}\|_{\mathrm{op}}$ converges to 0 in probability. Then $\nu_{D_n A_n}$ converges weakly in probability to ν.*

Proof. Statement (i) follows after a Hermitian dilation (12) from [8, Corollary A.41], see also [73, Exercise 2.4.3]. Statement (ii) follows after a Hermitian dilation from [8, Theorem A.43], see also [73, Exercise 2.4.4]. Let us now provide a proof for the statement (iii).

Pick some arbitrarily small $\varepsilon > 0$. It has to be shown that

$$\mathbb{P}\big(|\textstyle\int f(x)\mathrm{d}\nu_{D_n A_n}(x) - \int f(x)\mathrm{d}\nu(x)| > \varepsilon\big) = o(1) \tag{58}$$

for every continuous bounded function $f \in \mathcal{C}_b(\mathbb{R})$. Since ν and $\nu_{D_n A_n}$ are probability measures we may further restrict ourselves to the case where f is compactly supported. This follows by consideration of fg with g a bump function; see e.g. [45, Proof of Lemma 6.21]. Let $c_f \in \mathbb{R}_{>0}$ be a sufficiently large constant so that $\mathrm{supp}(f) \subseteq [-c_f, c_f]$. By f being compactly supported and continuous it follows that f is uniformly continuous. There therefore exists some $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon/2$ whenever $|x - y| < \delta$.

Observe that $\|D_n\|_{\mathrm{op}} \leq 1 + \|D_n - \mathrm{Id}\|_{\mathrm{op}}$ and $\|D_n^{-1}\|_{\mathrm{op}}^{-1} \geq 1 - \|D_n - \mathrm{Id}\|_{\mathrm{op}}$. Further, it follows from [41, Theorem 3.3.16(d)] that $\sigma_i(D_n A_n) \leq \|D_n\|_{\mathrm{op}}\sigma_i(A_n)$ and $\sigma_i(D_n A_n) \geq \|D_n^{-1}\|_{\mathrm{op}}^{-1}\sigma_i(A_n)$. When combined this yields that $\sigma_i(D_n A_n) - \sigma_i(A_n) \leq \|D_n - \mathrm{Id}\|_{\mathrm{op}}\sigma_i(A_n)$ and $\sigma_i(D_n A_n) - \sigma_i(A_n) \geq -\|D_n - \mathrm{Id}\|_{\mathrm{op}}\sigma_i(A_n)$, respectively. This is equivalent to saying that

$$|\sigma_i(D_n A_n) - \sigma_i(A_n)| \leq \|D_n - \mathrm{Id}\|_{\mathrm{op}}\sigma_i(A_n). \tag{59}$$

Since $\|D_n - \mathrm{Id}\|_{\mathrm{op}}$ converges to zero in probability it holds that

$$\mathbb{P}\big(\|D_n - \mathrm{Id}\|_{\mathrm{op}} < \min\{\delta/2c_f, 1/2\}\big) = 1 - o(1). \tag{60}$$

Deduce from (59) that whenever the event contained on the left-hand side of (60) holds, it follows that $\sigma_i(D_n A_n) > c_f$ for any i with $\sigma_i(A_n) > 2c_f$ and that $|\sigma_i(D_n A_n) - \sigma_i(A_n)| < \delta$ for any i with $\sigma_i(A_n) \leq 2c_f$. Therefore, whenever the event contained on the left-hand side

of (60) holds, it follows that

$$
\left| \int f(x) \mathrm{d}\nu_{D_n A_n}(x) - \int f(x) \mathrm{d}\nu_{A_n}(x) \right| \le \frac{1}{n} \sum_{i:\sigma_i(A_n) \le 2c_f} |f(\sigma_i(D_n A_n)) - f(\sigma_i(A_n))|
$$

$$
< \frac{\varepsilon}{2}. \tag{61}
$$

We used here that $\mathrm{supp}(f) \subseteq [-c_f, c_f]$. Combine (60) with (61) and the assumption that ν_{A_n} converges to ν weakly in probability to conclude that (58) holds. This concludes the proof. □

We will ultimately use Lemma 6.7(iii) with $D_n := \hat{D}_X^{-1} \mathrm{diag}((\ell+1)\Pi_X)$ to replace $\hat{D}_X$ by a deterministic matrix. This requires that $\hat{D}_X \approx \mathrm{diag}((\ell+1)\Pi_X)$, which will be shown by means of a concentration inequality that follows from the short mixing times of block Markov chains.

Let Z be a Markov chain on the state space $\mathcal{V}$ which is irreducible and acyclic. For any $\varepsilon \in [0, 1)$ the ε-*mixing time* of Z is defined as $t_{\mathrm{mix}}^Z(\varepsilon) := \min\{t \in \mathbb{Z}_{\ge 0} : d(t) \le \varepsilon\}$ where

$$
d(t) := \sup_{z \in \mathcal{V}} d_{\mathrm{TV}}(\mathbb{P}(Z_t = - \mid Z_0 = z), \Pi_Z) \tag{62}
$$

and d_{TV} denotes the total variation distance defined in (11). Set $t_{\mathrm{mix}}^Z := t_{\mathrm{mix}}^Z(1/4)$. Observe that since X is a block Markov chain it holds that $t_{\mathrm{mix}}^X \le \max\{t_{\mathrm{mix}}^{\Sigma_X}, 1\}$ where $\Sigma_{X,t} := \sigma(X_t)$ is the induced Markov chain on the clusters $\{1, \ldots, K\}$. Observe furthermore that the dynamics of Σ_X are independent of n by definition of a block Markov chain, so that $t_{\mathrm{mix}}^{\Sigma_X}$ is a constant. Thus $t_{\mathrm{mix}}^X = O(1)$.

We will rely on a concentration inequality from [59] which we reproduce here for the reader's convenience. Similar proofs for concentration in block Markov chains using this concentration inequality may be found in [66,67].

The concentration inequality is provided in terms of an invariant γ_{ps}^Z called the *pseudo-spectral gap*:

$$
\gamma_{ps}^Z := \max_{i \in \mathbb{Z}_{\ge 1}} \frac{1 - \lambda_2((P_Z^*)^i P_Z^i)}{i} \quad \text{where } P_Z^* := \mathrm{diag}(\Pi_Z)^{-1} P_Z^{\mathrm{T}} \mathrm{diag}(\Pi_Z). \tag{63}
$$

Here P_Z denotes the transition matrix of the Markov chain Z and P_Z^{T} is the transpose of this matrix. The pseudo-spectral gap is closely related to the mixing time:

$$
\frac{1}{2t_{\mathrm{mix}}^Z} \le \gamma_{ps}^Z \le \frac{1 + 2\ln(2) + \ln(1/\min_{i \in \mathcal{V}} \Pi_Z(i))}{t_{\mathrm{mix}}^Z} \tag{64}
$$

by [59, Proposition 3.4] whose assumptions are satisfied because Z is trivially uniformly ergodic as an irreducible and acyclic chain on a finite state space.

For any function $f : \mathcal{V} \to \mathbb{R}$ denote $\mathbb{E}_{\Pi_Z}[f]$ and $\mathrm{Var}_{\Pi_Z}[f]$ for the expectation and variance of the random variable $f(Z_0)$, respectively, where Z_0 has distribution Π_Z. The following proposition occurs in [59] in a more general setting with possibly infinite state spaces. We state the result here for the reader's convenience since it will be used multiple times.

Proposition 6.8 (*[59, Theorem 3.4]*). *Let Z be an irreducible and acyclic Markov chain on $\mathcal{V}$ starting in equilibrium. Consider a function $f : \mathcal{V} \to \mathbb{R}$ on the state space such that $|f(z) - \mathbb{E}_{\Pi_Z}[f]| \leq C$ for all $z \in \mathcal{V}$. Then, with $S := \sum_{t=0}^{\ell} f(Z_t)$, it holds that*

$$\mathbb{P}(|S - \mathbb{E}[S]| > r) \leq 2 \exp\left(-\frac{r^2 \gamma_{\text{ps}}^Z}{8(\ell + 1 + 1/\gamma_{\text{ps}}^Z)\operatorname{Var}_{\Pi_Z}[f] + 20rC}\right) \tag{65}$$

for all $r \in \mathbb{R}_{\geq 0}$.

Lemma 6.9. *Let $X := (X_t)_{t=0}^{\ell}$ be a sample path of the block Markov chain with an arbitrary initial distribution. Then, there exist constants $c_4, c_5 \in \mathbb{R}_{>0}$ not depending on the initial distribution such that*

$$\mathbb{P}(|\hat{D}_{X,ii} - (\ell + 1)\Pi_X(i)| > \sqrt{n}t) \leq c_4 \exp(-c_5 t) \tag{66}$$

for all $i = 1, \ldots, n$.

Proof. Let $\kappa \in \{1, \ldots, K\}$ be a random variable following π which is independent of X. Define

$$T := \inf\{t \in \mathbb{Z}_{\geq 1} : \sigma(X_t) = \kappa\}. \tag{67}$$

Observe that since the Markov chain associated with p is assumed to be irreducible and acyclic, it follows that $\mathbb{P}(T > t) \leq c_6 \exp(-c_7 t)$ for some $c_6, c_7 \in \mathbb{R}_{>0}$ which are independent of n and the initial state X_0.

Write $S_1 := \sum_{t=0}^{T-1} f(X_t) - \sum_{t=\ell+1}^{\ell+T} f(X_t)$ and $S_2 := \sum_{t=T}^{\ell+T} f(X_t)$ where $f(x) = \mathbb{1}_{x=i}$. We may apply Proposition 6.8 to derive a concentration inequality for S_2. Observe that

$$|f(x) - \mathbb{E}_{\Pi_X}[f]| \leq 1 \quad \text{and} \quad \operatorname{Var}_{\Pi_X}[f] = \Pi_X(i)(1 - \Pi_X(i)) \leq c_8 n^{-1} \tag{68}$$

for some $c_8 \in \mathbb{R}_{>0}$. Recall that $t_{\text{mix}}^X = O(1)$. Correspondingly, by (64), it follows that there exists some constant $c_9 \in \mathbb{R}_{>0}$ such that $\gamma_{\text{ps}}^X \geq c_9$. By Proposition 6.8 and the fact that $\ell = \Theta(n^2)$, we find that for any $r \in \mathbb{R}_{>0}$

$$\mathbb{P}(|S_2 - (\ell + 1)\Pi_X(i)| > \sqrt{n}r)$$
$$\leq 2 \exp\left(-\frac{nr^2 \gamma_{\text{ps}}^X}{8(\ell + 1 + 1/\gamma_{\text{ps}}^X)\operatorname{Var}_{\Pi_X}[f] + 20\sqrt{n}r}\right) \tag{69}$$
$$\leq c_{10} \exp(-c_{11} r). \tag{70}$$

The constant $c_{11} \in \mathbb{R}_{>0}$ is chosen so that $c_{11} \leq n\gamma_{\text{ps}}^X/(8(\ell + 1 + 1/\gamma_{\text{ps}}^X)\operatorname{Var}_{\Pi_X}[f] + 20\sqrt{n})$ for all n and the constant $c_{10} \geq 2$ is chosen so that $c_{10} \exp(-c_{11}r) \geq 1$ whenever $r \leq 1$.

Further, since T has exponential decay and $f \leq 1$, there exist constants $c_{12}, c_{13} \in \mathbb{R}_{>0}$ such that

$$\mathbb{P}(S_1 > \sqrt{n}r) \leq \mathbb{P}(T \geq \sqrt{n}r/2) \leq c_{12} \exp(-c_{13}\sqrt{n}r). \tag{71}$$

The desired result now follows by the triangle inequality and the fact that $|\hat{D}_{X,ii} - \sum_{i=0}^{\ell} f(X_t)| \leq 1$. $\square$

The following corollary is immediate by the union bound and the fact that $(\ell + 1)\Pi_X$ has entries of size $\Theta(n)$.

Corollary 6.10. *Let $X := (X_t)_{t=0}^\ell$ be a sample path from the block Markov chain with an arbitrary initial distribution. Then $\| \operatorname{diag}((\ell + 1)\Pi_X)^{-1}\hat{D}_X - \operatorname{Id} \|_{\mathrm{op}}$ converges to zero almost surely as n tends to infinity. In particular, asymptotically, $\hat{D}_X$ is invertible almost surely and $\hat{P}_X$ thus well-defined.*

Let us now proceed to the proof of Lemma 4.4. Recall that Lemma 4.4 concerns a reduction to centered random matrices when starting the chain from equilibrium.

Proof of Lemma 4.4. Let $\widetilde{X} := (\widetilde{X}_t)_{t=0}^\infty$ and $\widetilde{Y} := (\widetilde{Y}_t)_{t=0}^\infty$ denote infinitely long sample paths of the block Markov chain where $\widetilde{X}$ starts in equilibrium and $\widetilde{Y}$ has some arbitrary initial distribution. Define

$$T := \inf\{t \in \mathbb{Z}_{\geq 1} : \sigma(\widetilde{X}_t) = \sigma(\widetilde{Y}_t)\} \tag{72}$$

and observe that since the Markov chain associated with p is irreducible and acyclic, it holds that $\mathbb{P}(T > t) \leq c_{14} \exp(-c_{15}t)$ for some constants $c_{14}, c_{15} \in \mathbb{R}_{>0}$ which are independent of n. Let $X := (X_t)_{t=0}^\ell$ be the path defined by $X_t = \widetilde{X}_t$ for all $t < T$ and $X_t = \widetilde{Y}_t$ for all $t \geq T$. Further, let $Y := (\widetilde{Y}_t)_{t=0}^\ell$ be the truncation of $\widetilde{Y}$ to length $\ell + 1$. Observe that X is a sample path from the block Markov chain starting in equilibrium whereas Y is a sample path from the block Markov chain with an arbitrary initial distribution. The desired results concern the singular value distributions associated with Y.

Set

$$B_n := \frac{\hat{N}_Y - \hat{N}_X}{\sqrt{n}} \quad \text{and} \quad C_n := \frac{\mathbb{E}[\hat{N}_X]}{\sqrt{n}}, \tag{73}$$

and observe that $\hat{N}_Y/\sqrt{n} = M_X/\sqrt{n} + B_n + C_n$ almost surely. The first claim now follows by Lemma 6.7 since $\|B_n\|_{\mathrm{F}}^2/n \leq (2T/\sqrt{n})^2/n$ converges to zero in probability and $\operatorname{rank}(C_n) \leq K$.

For the second claim set $D_n := \hat{D}_Y^{-1}\operatorname{diag}((\ell + 1)\Pi_X)$ and observe that $\sqrt{n}\hat{P}_Y = D_n\sqrt{n}\operatorname{diag}((\ell + 1)\Pi_X)^{-1}\hat{N}_Y$. By Corollary 6.10 and the continuity of $M \mapsto M^{-1}$ in the neighborhood of Id it holds that $\|D_n - \operatorname{Id}\|_{\mathrm{op}}$ converges to zero in probability. Hence, it is sufficient to establish that the singular value distribution of $\sqrt{n}\operatorname{diag}((\ell + 1)\Pi_X)^{-1}\hat{N}_Y$ has the desired weak limit in probability by Lemma 6.7(iii).

In this regard observe that, with notation as in (73),

$$\sqrt{n}\operatorname{diag}((\ell + 1)\Pi_X)^{-1}\hat{N}_Y \tag{74}$$
$$= \sqrt{n}Q_X + n\operatorname{diag}((\ell + 1)\Pi_X)^{-1}B_n + n\operatorname{diag}((\ell + 1)\Pi_X)^{-1}C_n.$$

Here, it holds that

$$\|n\operatorname{diag}((\ell + 1)\Pi_X)^{-1}B_n\|_{\mathrm{F}}^2/n \leq \|n\operatorname{diag}((\ell + 1)\Pi_X)^{-1}\|_{\mathrm{op}}^2\|B_n\|_{\mathrm{F}}^2/n. \tag{75}$$

Hence, since we already know that $\|B_n\|_{\mathrm{F}}^2/n$ converges to zero in probability and since $\|n\operatorname{diag}((\ell + 1)\Pi_X)^{-1}\|_{\mathrm{op}} = \Theta(1)$ it follows that $\|n\operatorname{diag}((\ell + 1)\Pi_X)^{-1}B_n\|_{\mathrm{F}}^2/n$ converges to zero in probability. Furthermore,

$$\operatorname{rank}(n\operatorname{diag}((\ell + 1)\Pi_X)^{-1}C_n) \leq \operatorname{rank}(C_n) \leq K. \tag{76}$$

Apply Lemma 6.7 to (74) to conclude the proof. $\square$

6.3.2. Proof of the Poisson limit Theorem 4.5

We will extract Theorem 4.5 from a nonasymptotic result. For such a nonasymptotic result one has to precisely quantify the mixing behavior of the chain of clusters $\Sigma_X = (\sigma(X_t))_{t=0}^{\ell}$. For our purposes this is most naturally done in terms of the relative pointwise distance.

Let Z be a Markov chain on the state space $\mathcal{V}$ which is irreducible and acyclic. The *relative pointwise distance* $\Delta_Z(r)$ after $r \in \mathbb{Z}_{\geq 1}$ steps is given by

$$\Delta_Z(r) := \max_{x,y \in \mathcal{V}} \frac{|\mathbb{P}(Z_r = y \mid Z_0 = x) - \Pi_Z(y)|}{\Pi_Z(y)}. \tag{77}$$

Note that $\Delta_Z(r)$ is related to the quantity $d(r)$ from (62) which was used to define the mixing time. Indeed, a direct calculation with the definitions shows that

$$0 \leq \Delta_Z(r) \leq \frac{\max_{x,y \in \mathcal{V}} |\mathbb{P}(Z_r = y \mid Z_0 = x) - \Pi_Z(y)|}{\min_{y \in \mathcal{V}} \Pi_Z(y)} = (\min_{y \in \mathcal{V}} \Pi_Z(y))^{-1} d(r). \tag{78}$$

Theorem 6.11. *Let $X = (X_t)_{t=0}^{\ell}$ be a sample path from a block Markov chain which starts in equilibrium. Pick some $\varepsilon \in [0, 1/2]$ and $r_0 \in \mathbb{Z}_{\geq 1}$ such that $\Delta_{\Sigma_X}(r) \leq \varepsilon$ for all $r \geq r_0$. Then, for any $k_1, k_2 \in \{1, \ldots, K\}$ and $e \in \mathcal{V}_{k_1} \times \mathcal{V}_{k_2}$ which is not a self-loop*

$$d_{\mathrm{TV}}\left(\mathbb{P}(\hat{N}_{X,e} = -), \mathrm{Poisson}\left(\frac{\ell \pi(k_1) p_{k_1,k_2}}{\#\mathcal{V}_{k_1} \#\mathcal{V}_{k_2}}\right)\right) \tag{79}$$

$$\leq \ell \pi(k_1) p_{k_1,k_2}\left(\frac{4r_0 - 1}{(\#\mathcal{V}_{k_1})^2 (\#\mathcal{V}_{k_2})^2} + \frac{12\varepsilon}{\#\mathcal{V}_{k_1} \#\mathcal{V}_{k_2}}\right),$$

and for any self-loop $e \in \mathcal{V}_{k_1} \times \mathcal{V}_{k_1}$

$$d_{\mathrm{TV}}\left(\mathbb{P}(\hat{N}_{X,e} = -), \mathrm{Poisson}\left(\frac{\ell \pi(k_1) p_{k_1,k_1}}{(\#\mathcal{V}_{k_1})^2}\right)\right) \tag{80}$$

$$\leq \ell \pi(k_1) p_{k_1,k_1}\left(\frac{4r_0 - 1}{(\#\mathcal{V}_{k_1})^4} + \frac{2}{(\#\mathcal{V}_{k_1})^3} + \frac{12\varepsilon}{(\#\mathcal{V}_{k_1})^2}\right).$$

Proof. The proof consists of the following parts. Part 1 introduces notation which is used in [6] to quantify local and long-range dependence. The parameters quantifying the local dependence are estimated in Part 2. Finally, the parameter quantifying long-range dependence is estimated in Part 3.

Part 1: Notation for local and long-range dependence

For every $t \in \{1, \ldots, \ell\}$ let $B_{r_0}(t) := \{t' \in \{1, \ldots, \ell\} : |t' - t| \leq r_0\}$ and consider the following parameters which quantify local and long-range dependencies:

$$\begin{aligned}
\mathfrak{p}_t &:= \mathbb{P}(E_{X,t} = e), \\
\mathfrak{p}_{t,t'} &:= \mathbb{P}(E_{X,t} = e, E_{X,t'} = e), \\
\mathfrak{s}_t &:= \mathbb{E}\left[\left|\mathbb{P}\left(E_{X,t} = e \mid (\mathbb{1}_{E_{X,t'}=e})_{t' \in \{1,\ldots,\ell\} \setminus B_{r_0}(t)}\right) - \mathfrak{p}_t\right|\right].
\end{aligned} \tag{81}$$

Applying [6, Theorem 1] to the sum of dependent random variables $\hat{N}_{X,e} = \sum_{t=1}^{\ell} \mathbb{1}_{E_{X,t}=e}$ yields that

$$d_{\mathrm{TV}}\left(\mathbb{P}(\hat{N}_{X,e} = -), \mathrm{Poisson}\left(\frac{\ell \pi(k_1) p_{k_1,k_2}}{\#\mathcal{V}_{k_1} \#\mathcal{V}_{k_2}}\right)\right) \leq b_1 + b_2 + b_3. \tag{82}$$

where

$$b_1 := \sum_{t=1}^{\ell} \sum_{t' \in B_{r_0}(t)} \mathfrak{p}_t \mathfrak{p}_{t'}, \quad b_2 := \sum_{t=1}^{\ell} \sum_{t' \in B_{r_0}(t) \setminus \{t\}} \mathfrak{p}_{t,t'}, \quad b_3 := \sum_{t=1}^{\ell} \mathfrak{s}_t. \tag{83}$$

Here we used that the definition in [6] for total variation distance differs from our definition (11) by a factor two. Indeed, observe that $\sup_{A \subseteq \mathbb{Z}_{\geq 0}} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{x \in \mathbb{Z}_{\geq 0}} |\mu(x) - \nu(x)|$ for any two probability measures μ and ν using that the supremum is realized by $A = \{x \in \mathbb{Z}_{\geq 0} : \mu(x) \geq \nu(x)\}$.

Part 2: Bounding the local dependence contributions (b_1 and b_2)

By definition of X as a block Markov chain it holds that for any $t \in \{1, \ldots, \ell\}$

$$\mathfrak{p}_t = \frac{\mathbb{P}(\Sigma_{X,t-1} = k_1, \Sigma_{X,t} = k_2)}{\#\mathcal{V}_{k_1} \#\mathcal{V}_{k_2}} = \frac{\pi(k_1) p_{k_1,k_2}}{\#\mathcal{V}_{k_1} \#\mathcal{V}_{k_2}}. \tag{84}$$

From (84) it follows that

$$b_1 \leq \ell(2r_0 + 1) \frac{\pi(k_1)^2 p_{k_1,k_2}^2}{(\#\mathcal{V}_{k_1})^2 (\#\mathcal{V}_{k_2})^2} \leq \ell \pi(k_1) p_{k_1,k_2} \frac{2r_0 + 1}{(\#\mathcal{V}_{k_1})^2 (\#\mathcal{V}_{k_2})^2} \tag{85}$$

where it was used that $\pi(k_1) p_{k_1,k_2} \leq 1$. It holds for any $t, t' \in \{1, \ldots, \ell\}$ with $|t - t'| > 1$ that

$$\mathfrak{p}_{t,t'} = \frac{\mathbb{P}(\Sigma_{X,t-1} = k_1, \Sigma_{X,t} = k_2, \Sigma_{X,t'-1} = k_1, \Sigma_{X,t'} = k_2)}{(\#\mathcal{V}_{k_1})^2 (\#\mathcal{V}_{k_2})^2} \tag{86}$$

$$\leq \frac{\pi(k_1) p_{k_1,k_2}}{(\#\mathcal{V}_{k_1})^2 (\#\mathcal{V}_{k_2})^2}. \tag{87}$$

For $t' \in \{t + 1, t - 1\}$ it may similarly be deduced that if e is a self-loop then $\mathfrak{p}_{t,t'} \leq \pi(k_1) p_{k_1,k_1} (\#\mathcal{V}_{k_1})^{-3}$ and if e is not a self-loop then $\mathfrak{p}_{t,t'} = 0$. Hence,

$$b_2 \leq \begin{cases} \ell \pi(k_1) p_{k_1,k_1} \left(\frac{2}{(\#\mathcal{V}_{k_1})^3} + \frac{2r_0 - 2}{(\#\mathcal{V}_{k_1})^4} \right) & \text{if } e \text{ is a self-loop,} \\ \ell \pi(k_1) p_{k_1,k_2} \frac{2r_0 - 2}{(\#\mathcal{V}_{k_1})^2 (\#\mathcal{V}_{k_2})^2} & \text{otherwise.} \end{cases} \tag{88}$$

Part 3: Bounding the long-range dependence contribution (b_3)

The first step in this part will be to control $\mathfrak{s}_t$ in terms of a simpler quantity; this step is achieved in (92). Thereafter, control over (92) is achieved in (101) and we deduce the desired bound for b_3 in (105).

Recall that X is uniformly distributed in $\prod_{t=0}^{\ell} \mathcal{V}_{s_{X,t}}$ given that $\Sigma_X = s_X$. It follows that $\mathbb{1}_{E_{X,t}=e}$ is conditionally independent of $(\mathbb{1}_{E_{X,t}=e})_{t' \in \{1,\ldots,\ell\} \setminus B_{r_0}(t)}$ given $\Sigma_{X,t-r_0-1}$ and $\Sigma_{X,t+r_0}$. (If $t - r_0 - 1 < 0$ one can use a time reversal to make sense of $\Sigma_{X,t-r_0-1}$; a time reversal exists since Σ_X is assumed to be irreducible and acyclic.) Hence, by the law of total probability

$$\mathbb{P}(E_{X,t} = e \mid (\mathbb{1}_{E_{X,t}=e})_{t' \in \{1,\ldots,\ell\} \setminus B_{r_0}(t)}) \tag{89}$$

$$= \sum_{s_1, s_2 \in \{1,\ldots,K\}} \mathbb{P}(\Sigma_{X,t-r_0-1} = s_1, \Sigma_{X,t+r_0} = s_2 \mid (\mathbb{1}_{E_{X,t}=e})_{t' \in \{1,\ldots,\ell\} \setminus B_{r_0}(t)})$$

$$\times \mathbb{P}(E_{X,t} = e \mid \Sigma_{X,t-r_0-1} = s_1, \Sigma_{X,t+r_0} = s_2).$$

On the other hand, we may write

$$
\mathfrak{p}_t \tag{90}
$$
$$
= \mathfrak{p}_t \sum_{s_1, s_2 \in \{1, \ldots, K\}} \mathbb{P}\big(\Sigma_{X,t-r_0-1} = s_1, \Sigma_{X,t+r_0} = s_2 \mid (\mathbb{1}_{E_{X,t}=e})_{t' \in \{1,\ldots,\ell\} \setminus B_{r_0}(t)}\big).
$$

Observe that

$$
\mathbb{E}\big[\mathbb{P}(\Sigma_{X,t-r_0-1} = s_1, \Sigma_{X,t+r_0} = s_2 \mid (\mathbb{1}_{E_{X,t}=e})_{t' \in \{1,\ldots,\ell\} \setminus B_{r_0}(t)})\big] \tag{91}
$$
$$
= \mathbb{P}(\Sigma_{X,t-r_0-1} = s_1, \Sigma_{X,t+r_0} = s_2).
$$

Therefore, by using the triangle inequality on $\mathfrak{s}_t$'s definition together with (89)–(91), it follows that

$$
\mathfrak{s}_t \leq \sum_{s_1, s_2 \in \{1, \ldots, K\}} \mathbb{P}(\Sigma_{X,t-r_0-1} = s_1, \Sigma_{X,t+r_0} = s_2) \tag{92}
$$
$$
\times \big|\mathbb{P}(E_{X,t} = e \mid \Sigma_{X,t-r_0-1} = s_1, \Sigma_{X,t+r_0} = s_2) - \mathfrak{p}_t\big|.
$$

Observe that for any $s_1, s_2 \in \{1, \ldots, K\}$ with $\mathbb{P}(\Sigma_{X,t+r_0} = s_2 \mid \Sigma_{X,t-r_0-1} = s_1) > 0$, the definition of conditional probability yields that

$$
\mathbb{P}(E_{X,t} = e \mid \Sigma_{X,t-r_0-1} = s_1, \Sigma_{X,t+r_0} = s_2) \tag{93}
$$
$$
= \frac{\mathbb{P}(E_{X,t} = e, \Sigma_{X,t+r_0} = s_2 \mid \Sigma_{X,t-r_0-1} = s_1)}{\mathbb{P}(\Sigma_{X,t+r_0} = s_2 \mid \Sigma_{X,t-r_0-1} = s_1)}.
$$

Here, by definition of a block Markov chain

$$
\mathbb{P}(E_{X,t} = e, \Sigma_{X,t+r_0} = s_2 \mid \Sigma_{X,t-r_0-1} = s_1) \tag{94}
$$
$$
= \frac{\mathbb{P}(\Sigma_{X,t-1} = k_1, \Sigma_{X,t} = k_2, \Sigma_{X,t+r_0} = s_2 \mid \Sigma_{X,t-r_0-1} = s_1)}{\#\mathcal{V}_{k_1} \#\mathcal{V}_{k_2}}.
$$

Moreover, the Markov property applied to Σ_X yields that

$$
\mathbb{P}(\Sigma_{X,t-1} = k_1, \Sigma_{X,t} = k_2, \Sigma_{X,t+r_0} = s_2 \mid \Sigma_{X,t-r_0-1} = s_1) \tag{95}
$$
$$
= \mathbb{P}(\Sigma_{X_{t-1}} = k_1 \mid \Sigma_{X,t-r_0-1} = s_1) p_{k_1,k_2} \mathbb{P}(\Sigma_{X,t+r_0} = s_2 \mid \Sigma_{X,t} = k_2).
$$

Combining (93), (94) and (95) yields that

$$
\mathbb{P}(E_{X,t} = e \mid \Sigma_{X,t-r_0-1} = s_1, \Sigma_{X,t+r_0} = s_2) \tag{96}
$$
$$
= \frac{p_{k_1,k_2}}{\#\mathcal{V}_{k_1} \#\mathcal{V}_{k_2}} \frac{\mathbb{P}(\Sigma_{X,t-1} = k_1 \mid \Sigma_{X,t-r_0-1} = s_1) \mathbb{P}(\Sigma_{X,t+r_0} = s_2 \mid \Sigma_{X,t} = k_2)}{\mathbb{P}(\Sigma_{X,t+r_0} = s_2 \mid \Sigma_{X,t-r_0-1} = s_1)}.
$$

Let us introduce the following notation:

$$
\delta^{(a)}_{s_1,s_2} := \frac{\mathbb{P}(\Sigma_{X,t+r_0} = s_2 \mid \Sigma_{X,t-r_0-1} = s_1) - \pi(s_2)}{\pi(s_2)}, \tag{97}
$$
$$
\delta^{(b)}_{s_1,k_1} := \frac{\mathbb{P}(\Sigma_{X,t-1} = k_1 \mid \Sigma_{X,t-r_0-1} = s_1) - \pi(k_1)}{\pi(k_1)}, \tag{98}
$$
$$
\delta^{(c)}_{k_2,s_2} := \frac{\mathbb{P}(\Sigma_{X,t+r_0} = s_2 \mid \Sigma_{X,t} = k_2) - \pi(s_2)}{\pi(s_2)}. \tag{99}
$$

Substituting (97)–(99) into (96) yields

$$\mathbb{P}(E_{X,t} = e \mid \Sigma_{X,t-r_0-1} = s_1, \Sigma_{X,t+r_0} = s_2) \tag{100}$$

$$= \frac{\pi(k_1)p_{k_1,k_2}}{\#\mathcal{V}_{k_1}\#\mathcal{V}_{k_2}} \frac{(1 + \delta^{(b)}_{s_1,k_1})(1 + \delta^{(c)}_{k_2,s_2})}{1 + \delta^{(a)}_{s_1,s_2}}. \tag{101}$$

Note that by our assumption and the definition of relative pointwise distance, $|\delta^{(a)}_{s_1,s_2}| \leq \varepsilon \leq 1/2$. Hence by Taylor's theorem

$$\left| 1 - \frac{1}{1 + \delta^{(a)}_{s_1,s_2}} \right| \leq \varepsilon \max_{x \in [-1/2, 1/2]} \left(\frac{1}{1 + x} \right)^2 \leq 4\varepsilon. \tag{102}$$

Similarly, it holds that $|\delta^{(b)}_{s_1,k_1}| \leq \varepsilon \leq 1/2$ and $|\delta^{(c)}_{k_2,s_2}| \leq \varepsilon \leq 1/2$. In particular $|(1 + \delta^{(b)}_{s_1,k_1})(1 + \delta^{(c)}_{k_2,s_2})| \leq 9/4$ and

$$|1 - (1 + \delta^{(b)}_{s_1,k_1})(1 + \delta^{(c)}_{k_2,s_2})| \leq 3\varepsilon. \tag{103}$$

Bound (101) using (84), (102), (103) to find that

$$|\mathbb{P}(E_{X,t} = e \mid \Sigma_{X,t-r_0-1} = s_1, \Sigma_{X,t+r_0} = s_2) - \mathfrak{p}_t| \leq \frac{\pi(k_1)p_{k_1,k_2}}{\#\mathcal{V}_{k_1}\#\mathcal{V}_{k_2}} \cdot 12\varepsilon. \tag{104}$$

Hence by (92) and the definition of b_3,

$$b_3 \leq \ell\pi(k_1)p_{k_1,k_2} \frac{12\varepsilon}{\#\mathcal{V}_{k_1}\#\mathcal{V}_{k_2}}. \tag{105}$$

This concludes the proof: the estimates for b_1, b_2 and b_3 from (85), (88) and (105) can be substituted in (82) to yield the desired result. □

Proof of Theorem 4.5. Note that the total variation distance in (11) metrizes convergence in distribution. This is to say that $\hat{N}_{X,e_n}$ converges in distribution if and only if $\mathbb{P}(\hat{N}_{X,e_n} = -)$ converges with respect to the metric d_{TV}.

Recall that Σ_X is assumed to be irreducible and acyclic. Hence, by (78) it holds that $\Delta_{\Sigma_X}(r)$ decays exponentially in r. Recall that $\ell = \lambda n^2 + o(n^2)$ and $\#\mathcal{V}_k = \alpha_k n + o(n)$. In particular $(\#\mathcal{V}_{k_1}\#\mathcal{V}_{k_2})^{-1}\ell$ converges to $\alpha_{k_1}^{-1}\alpha_{k_2}^{-1}\lambda$ as n tends to infinity. The result now follows by taking $r_0 = \log(\#\mathcal{V}_{k_1}\#\mathcal{V}_{k_2})$ in Theorem 6.11. □

6.3.3. *Proof of Corollary 4.6 on the variance profile*

Due to the decomposition $\text{Var}[\hat{N}_{X,e_n}] = \mathbb{E}[\hat{N}^2_{X,e_n}] - \mathbb{E}[\hat{N}_{X,e_n}]^2$ the following result implies Corollary 4.6.

Lemma 6.12. *Assume that X starts in equilibrium. Let e_n be as in Theorem 4.5, and Y a Poisson distributed random variable with rate $\lambda\pi(k_1)\alpha_{k_1}^{-1}\alpha_{k_2}^{-1}p_{k_1,k_2}$. Then for every $m \in \mathbb{Z}_{\geq 0}$ it holds that, as n tends to infinity,*

$$\mathbb{E}\left[\hat{N}^m_{X,e_n}\right] = \mathbb{E}\left[Y^m\right] + o_m(1). \tag{106}$$

Proof. We have already established in Theorem 4.5 that $\hat{N}_{X,e_n}$ converges in distribution to Y. Hence, to derive convergence of the moments it suffices to verify that for every $m \in \mathbb{Z}_{\geq 0}$ it holds that the sequence of random variables $\hat{N}^{m+1}_{X,e_n}$ is uniformly integrable [36, 7.10. (15)].

We will apply Proposition 6.8 to the Markov chain of edges $E_X := (E_{X,t})_{t=1}^{\ell}$ with the function $f(e) = \mathbb{1}_{e=e_n}$. Observe that

$$|f - \mathbb{E}_{\Pi_{E_X}}[f]| \le 1; \qquad \mathrm{Var}_{\Pi_{E_X}}[f] = \Pi_{E_X}(e_n)(1 - \Pi_{E_X}(e_n)) \le \alpha_{min}^{-2} n^{-2}. \tag{107}$$

Note that E_X induces a chain Σ_{E_X} on the reduced space $\{1, \ldots, K\} \times \{1, \ldots, K\}$ and that $t_{\mathrm{mix}}^{E_X} \le \max\{2, t_{\mathrm{mix}}^{\Sigma_{E_X}}\}$ by X being a block Markov chain. Since X is a block Markov chain it follows that the dynamics of Σ_{E_X} do not depend on n. It follows that $t_{\mathrm{mix}}^{\Sigma_{E_X}}$ is a constant. Correspondingly, by (64) it holds that $1/\gamma_{\mathrm{ps}}^{E_X} = O(1)$. Hence, since $\ell = \Theta(n^2)$, Proposition 6.8 yields that

$$\mathbb{P}(|\hat{N}_{X,e_n} - \ell \pi(k_1)(\#\mathcal{V}_{k_1})^{-1}(\#\mathcal{V}_{k_2})^{-1} p_{k_1,k_2}| > r)$$

$$\le 2 \exp\left(-\frac{r^2 \gamma_{\mathrm{ps}}^{E_X}}{8(\ell + 1/\gamma_{\mathrm{ps}}^{E_X})\mathrm{Var}_{\Pi_{E_X}}[f] + 20r}\right) \tag{108}$$

$$\le c_{17} \exp(-c_{16} r) \tag{109}$$

for some $c_{16}, c_{17} \in \mathbb{R}_{>0}$ which do not depend on n. The constant $c_{16} \in \mathbb{R}_{>0}$ is chosen so that $c_{16} \le \gamma_{\mathrm{ps}}^{E_X}/(8(\ell + 1/\gamma_{\mathrm{ps}}^{E_X})\mathrm{Var}_{\Pi_{E_X}}[f] + 20)$ for all n and the constant $c_{17} \ge 2$ is chosen so that $c_{17}\exp(-c_{16}r) \ge 1$ whenever $r \le 1$. Note that $\ell \pi(k_1)(\#\mathcal{V}_{k_1})^{-1}(\#\mathcal{V}_{k_2})^{-1} p_{k_1,k_2}$ converges to a finite constant as n tends to infinity. Hence, for some sufficiently large constant $c_{18} \in \mathbb{R}_{>0}$

$$\mathbb{P}(\hat{N}_{X,e_n} > r) \le c_{18}\exp(-c_{16}r). \tag{110}$$

In particular,

$$\mathbb{E}[\hat{N}_{X,e_n}^{m+1}] = \sum_{r=1}^{\infty} r^{m+1}\mathbb{P}(\hat{N}_{X,e_n} = r) \tag{111}$$

$$\le c_{18}\sum_{r=1}^{\infty} r^{m+1}\exp(-c_{16}(r-1)) \tag{112}$$

The right-hand side of (112) is finite and independent of n. This shows that $\hat{N}_{X,e_n}^{m+1}$ is uniformly integrable and concludes the proof. $\square$

Corollary 6.13. *Assume that X starts in equilibrium. Then, for all $m \in \mathbb{Z}_{\ge 0}$ it holds that as n tends to infinity $\max_{e \in \vec{E}_n} |\mathbb{E}[\hat{N}_{X,e}^m]| = O_m(1)$.*

Proof. Observe that $\mathbb{E}[\hat{N}_{X,e}^m]$ can only take K^2 possible values depending on the clusters of the endpoints of e. Hence, the result follows from Lemma 6.12. $\square$

6.3.4. Proof of Proposition 4.7 on $H(M_X)$ being approximately uncorrelated

Recall Definition 4.1 of an approximately uncorrelated random matrix with a variance profile. The definition consists of two statements: (14) and (15). These two parts are established in Propositions 6.15 and 6.14 respectively.

Proposition 6.14. *Assume that X starts in equilibrium and let $R \in \mathbb{Z}_{\ge 1}$. Then, for all $m \in \mathbb{Z}_{\ge 0}^R$ it holds that as n tends to infinity*

$$\max_{e_1, \ldots, e_R \in \vec{E}_n} \left| \mathbb{E}\left[M_{X,e_1}^{m_1} \cdots M_{X,e_R}^{m_R}\right] - \mathbb{E}\left[M_{X,e_1}^{m_1}\right] \cdots \mathbb{E}\left[M_{X,e_R}^{m_R}\right]\right| = o_{m,R}(1). \tag{113}$$
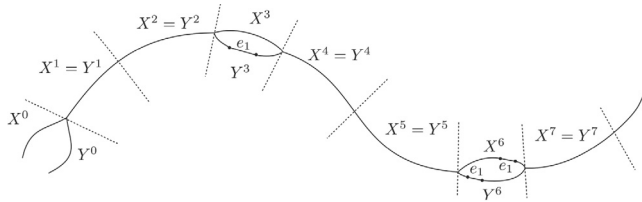
Fig. 7. Visualization of the merging process in the construction of X and Y during the proof of Proposition 6.14. The fragments X^i and Y^i are allowed to diverge whenever either one uses e_1. This ensures that all information about Y using e_1 is erased from X. Otherwise, the fragments are merged by taking $X^i = Y^i$. This ensures that the two chains are often equal because using e_1 is a rare event. The cluster structure is exploited to ensure that the endpoints of the short fragments X^i and Y^i can be glued after diverging.

Proof. We proceed by induction on R. The base-case $R = 1$ is trivial. Now assume that $R > 1$ and that the claim is known to hold for any product with strictly less than R edges e_j.

The proof is subdivided in three main steps. First, we construct a pair of Markov chains X, Y which are equal most of the time but are sometimes allowed to diverge. The goal of this process is to ensure two properties: (i) It holds that $M_{X,e_j} \approx M_{Y,e_j}$ for every $j \in \{2, \ldots, R\}$. (ii) It holds that X is independent of M_{Y,e_1}. The approximation from the first item may be used to show that

$$\mathbb{E}[M_{Y,e_1}^{m_1} M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}] \approx \mathbb{E}[M_{Y,e_1}^{m_1} M_{X,e_2}^{m_2} \cdots M_{X,e_R}^{m_R}]. \tag{114}$$

The independence from the second item allows us to factorize

$$\mathbb{E}[M_{Y,e_1}^{m_1} M_{X,e_2}^{m_2} \cdots M_{X,e_R}^{m_R}] = \mathbb{E}[M_{Y,e_1}^{m_1}]\mathbb{E}[M_{X,e_2}^{m_2} \cdots M_{X,e_R}^{m_R}]. \tag{115}$$

Combined, (114) and (115) yield that $\mathbb{E}[M_{Y,e_1}^{m_1} \cdots M_{Y,e_R}^{m_R}]$ equals $\mathbb{E}[M_{Y,e_1}^{m_1}]\mathbb{E}[M_{X,e_2}^{m_2} \cdots M_{X,e_R}^{m_R}]$ up to some error term. The induction hypothesis then yields the desired result provided we show that the error term is small. A bound on the size of the error terms is established in the third step.

Some preliminary reductions are applicable. Firstly, precisely as in Part 1 of the proof of Proposition 4.8, it can be assumed that $K \geq 2R+1$ by splitting a cluster into pieces of asymptotically equal proportions. Further, the value of $\mathbb{E}[M_{X,e_1}^{m_1} \cdots M_{X,e_R}^{m_R}] - \mathbb{E}[M_{X,e_1}^{m_1}] \cdots \mathbb{E}[M_{X,e_R}^{m_R}]$ can only depend on the isomorphism type of the labeled directed subgraph G induced by $\{e_1, \ldots, e_R\}$ with vertex-labels given by the clusters and edge-labels by the index j of e_j. As n tends to infinity the number of possible isomorphism types for G remains bounded. Hence, we can fix some isomorphism type for G. Informally, this allows us to pretend that the edges $e_1, \ldots, e_R \in \vec{E}_n$ stay fixed as n tends to infinity. Due to the induction hypothesis it may also be assumed that the e_j are distinct.

Part 1: Construction of Markov chains (X, Y)

Recall that it was ensured that $K \geq 2R+1$. Hence, there exists some $k \in \{1, \ldots, K\}$ such that $\mathcal{V}_k$ does not contain any endpoint of e_j for $j = 1, \ldots, R$.

Use the following procedure to construct a triple of sample paths $(\widetilde{X}, \widetilde{Y}, \widetilde{Z})$ with random length $\geq \ell + 1$. These will be trimmed to paths X, Y and Z of length exactly equal to $\ell + 1$ afterwards. See Fig. 7 for a visualization of the construction.

(i) Sample two independent infinite paths $(\widetilde{X}_t^0)_{t=0}^\infty$ and $(\widetilde{Z}_t^0)_{t=0}^\infty$ from the block Markov chain starting in equilibrium. Due to the assumption that the Markov chain associated with p

is irreducible and acyclic it holds that $T_0 := \inf\{t \geq 1 : \widetilde{X}_t^0 \in \mathcal{V}_k, \widetilde{Z}_t^0 \in \mathcal{V}_k\}$ is finite with probability one.

Set $(\widetilde{X}_t)_{t=0}^{T_0-1} := (\widetilde{X}_t^0)_{t=0}^{T_0-1}$ and $(\widetilde{Y}_t)_{t=0}^{T_0-1} := (\widetilde{Z}_t)_{t=0}^{T_0-1} := (\widetilde{Z}_t^0)_{t=0}^{T_0-1}$.

(ii) For $i = 1, \ldots, \ell$ extend $\widetilde{X}, \widetilde{Y}$ and $\widetilde{Z}$ by using the following procedure.

  (a) Sample two independent infinite paths $(\widetilde{X}_t^i)_{t=0}^\infty$ and $(\widetilde{Z}_t^i)_{t=0}^\infty$ from the block Markov chain with $\widetilde{X}_0^i = \widetilde{Z}_0^i$ chosen uniformly at random in $\mathcal{V}_k$.

  (b) Let $T_i := \inf\{t \in \mathbb{Z}_{\geq 1} : \widetilde{X}_t^i \in \mathcal{V}_k, \widetilde{Z}_t^i \in \mathcal{V}_k\}$. Due to the assumption that the Markov chain associated with p is irreducible and acyclic it will hold that T_i is finite with probability one.

  (c) If e_1 was traversed by $(\widetilde{X}_t^i)_{t=0}^{T_i-1}$ or $(\widetilde{Z}_t^i)_{t=0}^{T_i-1}$ then let $(\widetilde{Y}_t^i)_{t=0}^{T_i-1} := (\widetilde{Z}_t^i)_{t=0}^{T_i-1}$. Else, let $(\widetilde{Y}_t^i)_{t=0}^{T_i-1} := (\widetilde{X}_t^i)_{t=0}^{T_i-1}$.

  (d) Append $(\widetilde{X}_t^i)_{t=0}^{T_i-1}, (\widetilde{Y}_t^i)_{t=0}^{T_i-1}$ and $(\widetilde{Z}_t^i)_{t=0}^{T_i-1}$ to the previously constructed parts of $\widetilde{X}, \widetilde{Y}$ and $\widetilde{Z}$ respectively. This is to say that we define $\widetilde{X}_{t+\sum_{j=0}^{i-1} T_j} := \widetilde{X}_t^i, \widetilde{Y}_{t+\sum_{j=0}^{i-1} T_j} := \widetilde{Y}_t^i$ and $\widetilde{Z}_{t+\sum_{j=0}^{i-1} T_j} := \widetilde{Z}_t^i$ for $t = 0, \ldots, T_i - 1$.

The sampled paths $X^i := (\widetilde{X}_t^i)_{t=0}^{T_i-1}$, $Y^i := (\widetilde{Y}_t^i)_{t=0}^{T_i-1}$ and $Z^i := (\widetilde{Z}_t^i)_{t=0}^{T_i-1}$ used in the construction of $\widetilde{X}, \widetilde{Y}$ and $\widetilde{Z}$ will be called fragments. For any $i = 1, \ldots, \ell$ the ith fragments are said to be merged if $(\widetilde{X}_t^i)_{t=0}^{T_i-1}$ and $(\widetilde{Z}_t^i)_{t=0}^{T_i-1}$ did not traverse e_1. Denote $X := (\widetilde{X}_t)_{t=0}^\ell$, $Y := (\widetilde{Y}_t)_{t=0}^\ell$ and $Z := (\widetilde{Z}_t)_{t=0}^\ell$ for the truncations to length $\ell + 1$. Note that these are block Markov chains from the specified model.

We further claim that X is independent of M_{Y,e_1}. Let us remark that this is due to the special role of e_1 in the above construction; it will typically not be true that X is independent of $M_{Y,e'}$ for any other edge $e' \neq e_1$. Indeed, let $\widetilde{W}$ be the path found from $\widetilde{Z}$ by replacing $\widetilde{Z}_{\sum_{j=0}^i T_i}$ with a uniformly random node in $\mathcal{V}_k$ for every $i = 0, \ldots, \ell - 1$. Then X is independent of $W := (\widetilde{W}_t)_{t=0}^\ell$. In particular X will be independent of M_{W,e_1} since this is a function of W. Now note that M_{Y,e_1} is equal to M_{W,e_1} by definition of Y and the fact that $\mathcal{V}_k$ was chosen not to contain any of the endpoints of e_1. This establishes that X is independent of M_{Y,e_1} as desired.

Part 2: Approximate factorization of $\mathbb{E}[M_{Y,e_1}^{m_1} \cdots M_{Y,e_R}^{m_R}]$

For any $j = 2, \ldots, R$ define Δ_j to be the difference between the number of times e_j was used in Y and X, that is,

$$\Delta_j := \hat{N}_{Y,e_j} - \hat{N}_{X,e_j}. \tag{116}$$

Observe that $M_{Y,e_j} = M_{X,e_j} + \Delta_j$ for every $j \geq 2$ since $\mathbb{E}[\hat{N}_{Y,e_j}] = \mathbb{E}[\hat{N}_{X,e_j}]$; after all X and Y have the same distribution. Substitute the binomial expansion for every factor $M_{Y,e_j}^{m_j} = (M_{X,e_j} + \Delta_j)^{m_j}$ in $M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}$ and use the fact that M_{Y,e_1} is independent of the M_{X,e_j} with $j \geq 2$ to find that

$$\mathbb{E}[M_{Y,e_1}^{m_1} M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}] - \mathbb{E}[M_{Y,e_1}^{m_1}]\mathbb{E}[M_{X,e_2}^{m_2} \cdots M_{X,e_R}^{m_R}] \tag{117}$$

$$= \sum_{0 \leq m' \leq m} c_{m'} \mathbb{E}\left[M_{Y,e_1}^{m_1} \prod_{j=2}^R \Delta_j^{m'_j} \prod_{j=2}^R M_{X,e_j}^{m_j - m'_j} \right]$$

where the summation runs over vectors of integers of length R and the $c_{m'}$ are absolute constants with $c_{m'} = 0$ whenever $m'_j = 0$ for all $j \geq 2$. Note that $M_{X,e_2}^{m_2} \cdots M_{X,e_R}^{m_R}$ is a product with $R - 1$ factors. It follows that the induction hypothesis is applicable to $\mathbb{E}[M_{X,e_2}^{m_2} \cdots M_{X,e_R}^{m_R}]$

so that it remains to show that $\mathbb{E}[M_{Y,e_1}^{m_1} \prod_{j=2}^{R} \Delta_j^{m'_j} \prod_{j=2}^{R} M_{X,e_j}^{m_j - m'_j}] = o_{m,R}(1)$ whenever $m'_j \geq 1$ for some $j \geq 2$.

By the Cauchy–Schwarz inequality it holds that

$$\left| \mathbb{E}\left[M_{Y,e_1}^{m_1} \prod_{j=2}^{R} \Delta_j^{m'_j} \prod_{j=2}^{R} M_{X,e_j}^{m_j - m'_j} \right] \right| \leq \sqrt{\mathbb{E}\left[\prod_{j=2}^{R} \Delta_j^{2m'_j} \right] \mathbb{E}\left[M_{Y,e_1}^{2m_1} \prod_{j=2}^{R} M_{X,e_j}^{2(m_j - m'_j)} \right]}. \tag{118}$$

Due to the independence of X on M_{Y,e_1} we can write

$$\mathbb{E}\left[M_{Y,e_1}^{2m_1} \prod_{j=2}^{R} M_{X,e_j}^{2(m_j - m'_j)} \right] = \mathbb{E}\left[M_{Y,e_1}^{2m_1} \right] \mathbb{E}\left[\prod_{j=2}^{R} M_{X,e_j}^{2(m_j - m'_j)} \right] \tag{119}$$

which is $O_m(1)$ by [Corollary 6.13](#) and the induction hypothesis. It remains to show that $\mathbb{E}[\prod_{j=2}^{R} \Delta_j^{2m'_j}] = o_{m'}(1)$.

Define $\mathcal{M}_i$ to be the event where X^i and Y^i were merged and denote $\neg \mathcal{M}_i$ for the complement of this event. For any fixed $j \geq 2$ let $\Delta_{X,j} := \hat{N}_{X^0,e_j} + \sum_{i=1}^{\ell} \mathbb{1}_{\neg \mathcal{M}_i} \hat{N}_{X^i,e_j}$ denote the number of times when $\widetilde{X}$ used edge e_j in a fragment where X and Y were not merged and define $\Delta_{Y,j}$ similarly. Then

$$|\Delta_j| = |\hat{N}_{Y,e_j} - \hat{N}_{X,e_j}| \leq \Delta_{Y,j} + \Delta_{X,j}. \tag{120}$$

Substitute the definitions of $\Delta_{X,j}$ and $\Delta_{Y,j}$ in the resulting bound $\prod_{j=2}^{R} \Delta_j^{2m'_j} \leq \prod_{j=2}^{R} (\Delta_{Y,j} + \Delta_{X,j})^{2m'_j}$ and expand the monomial expression. Then, the Cauchy–Schwarz inequality reduces us to the statement that $\mathbb{E}[\hat{N}_{X^0,e_j}^{q}] = o_q(1)$ and $\mathbb{E}[(\sum_{i=1}^{\ell} \mathbb{1}_{\neg \mathcal{M}_i} \hat{N}_{X^i,e_j})^q] = o_q(1)$ for any $q \in \mathbb{Z}_{\geq 1}$ where it was used that Y has the same distribution as X. These claims are established in the next part of the proof. In fact, we will establish the stronger claims that $\mathbb{E}[\hat{N}_{X^i,e_j}^{q}] = O_q(n^{-2})$ for any $(i, j) \in \{0, 1\} \times \{1, \ldots, R\}$ and $\mathbb{E}[(\sum_{i=1}^{\ell} \mathbb{1}_{\neg \mathcal{M}_i} \hat{N}_{X^i,e_j})^q] = O_q(n^{-1})$ for any $i \geq 1$.

Part 3: Bounding $\mathbb{E}[N_{X^i,e_j}^{q}]$ and $\mathbb{E}[(\sum_{i=1}^{\ell} \mathbb{1}_{\neg \mathcal{M}_i} \hat{N}_{X^i,e_j})^q]$

Part 3.1: $\mathbb{E}[N_{X^i,e_j}^{q}] = O_q(n^{-2})$

Fix some $(i, j) \in \{0, 1\} \times \{1, \ldots, R\}$ and $q \in \mathbb{Z}_{\geq 1}$. By the law of total expectation

$$\mathbb{E}[\hat{N}_{X^i,e_j}^{q}] = \sum_{t=1}^{\infty} \mathbb{P}(T_i = t) \mathbb{E}[\hat{N}_{X^i,e_j}^{q} \mid T_i = t]. \tag{121}$$

Use $\hat{N}_{X^i,e_j} = \sum_{t=1}^{T_i} \mathbb{1}_{E_{X^i,t} = e_j}$ to derive the following bound

$$\mathbb{E}[\hat{N}_{X^i,e_j}^{q} \mid T_i = t] = \sum_{t_1,\ldots,t_q=1}^{t} \mathbb{P}(E_{X^i,t_1} = e_j, \ldots, E_{X^i,t_q} = e_j \mid T_i = t) \tag{122}$$

$$\leq \sum_{t_1,\ldots,t_q=1}^{t} \mathbb{P}(E_{X^i,t_1} = e_j \mid T_i = t). \tag{123}$$

For any edge e whose starting and ending point are in the same clusters as the starting and ending point of e_j it holds that $\mathbb{P}(E_{X^i,t_1} = e \mid T_i = t) = \mathbb{P}(E_{X^i,t_1} = e_j \mid T_i = t)$. There are at least $\alpha_{min}^2 n^2$ such edges so it follows that $\mathbb{P}(E_{X^i,t_1} = e_j \mid T_i = t) \leq \alpha_{min}^{-2} n^{-2}$.

(123) now yields that

$$\mathbb{E}[\hat{N}^q_{X^i,e_j} \mid T_i = t] \le \alpha^{-2}_{min} n^{-2} t^q. \tag{124}$$

By (121) and (124) it follows that

$$\mathbb{E}[\hat{N}^q_{X^i,e_j}] \le \alpha^{-2}_{min} n^{-2} \mathbb{E}[T^q_i]. \tag{125}$$

Here, $\mathbb{E}[T^q_i]$ is a finite constant which does not depend on n. Indeed, consider the product chain $\Sigma_{(\widetilde{X}^i, \widetilde{Z}^i)} := (\sigma(\widetilde{X}^i_t), \sigma(\widetilde{Z}^i_t))^\infty_{t=0}$ on the space of clusters $\{1, \ldots, K\} \times \{1, \ldots, K\}$. Then T_i is the first strictly positive time $\Sigma_{(\widetilde{X}^i, \widetilde{Z}^i)}$ is in (k, k). In particular, $\mathbb{E}[T^q_i]$ is independent of n. The fact that $\mathbb{E}[T^q_i]$ is finite is immediate from the fact that $\mathbb{P}(T_i > t)$ shows exponential decay in t since the Markov chain associated with p is assumed to be irreducible and acyclic. This concludes the proof of the statement that $\mathbb{E}[N^q_{X^i,e_j}] = O_q(n^{-2})$.

Part 3.2: $\mathbb{E}[(\sum^\ell_{i=1} \mathbb{1}_{\neg \mathcal{M}_i} \hat{N}_{X^i,e_j})^q] = O_q(n^{-1})$

Observe that for any $q \in \mathbb{Z}_{\ge 1}$

$$\mathbb{E}\left[\left(\sum^\ell_{i=1} \mathbb{1}_{\neg \mathcal{M}_i} \hat{N}_{X^i,e_j}\right)^q\right] = \sum^\ell_{i_1,\ldots,i_q=1} \mathbb{E}\left[\prod^q_{l=1} \mathbb{1}_{\neg \mathcal{M}_{i_l}} \hat{N}_{X^{i_l},e_j}\right] \tag{126}$$

$$= \sum^q_{d=1} \sum_{\#\{i_1,\ldots,i_q\}=d} \mathbb{E}\left[\prod^q_{l=1} \mathbb{1}_{\neg \mathcal{M}_{i_l}} \hat{N}_{X^{i_l},e_j}\right] \tag{127}$$

where the second sum in (127) runs over all values of $i_1, \ldots, i_q \in \{1, \ldots, \ell\}$ with $\#\{i_1, \ldots, i_q\} = d$.

Note that the random variables $\mathbb{1}_{\neg \mathcal{M}_i} \hat{N}_{X^i,e_j}$ with $i \ge 1$ are independent and identically distributed due to the fact that $\mathcal{V}_k$ was chosen not to contain any endpoint of e_1 or e_j. Correspondingly, a term with d distinct i_l values on the right-hand side of (127) may be factorized as $\prod^d_{l=1} \mathbb{E}[\mathbb{1}_{\neg \mathcal{M}_1} \hat{N}^{q_l}_{X^1,e_j}]$ for some $q_1, \ldots, q_l \in \{1, \ldots, q\}$. Since $\hat{N}_{X^1,e_j} \ge 1$ it holds that $\mathbb{E}[\mathbb{1}_{\neg \mathcal{M}_1} \hat{N}^{q_l}_{X^1,e_j}] \le \mathbb{E}[\mathbb{1}_{\neg \mathcal{M}_1} \hat{N}^q_{X^1,e_j}]$. Consequently, we may bound (127) as

$$\mathbb{E}\left[\left(\sum^\ell_{i=1} \mathbb{1}_{\neg \mathcal{M}_i} \hat{N}_{X^i,e_j}\right)^q\right] \le \sum^q_{d=1} c_{q,d} \ell^d \mathbb{E}\left[\mathbb{1}_{\neg \mathcal{M}_1} \hat{N}^q_{X^1,e_j}\right]^d \tag{128}$$

for some absolute constants $c_{q,d} \in \mathbb{R}_{>0}$. Since $\ell = \Theta(n^2)$ it now suffices to show that $\mathbb{E}[\mathbb{1}_{\neg \mathcal{M}_1} \hat{N}^q_{X^1,e_j}] = O_q(n^{-3})$. This will be shown by an argument resembling Part 3.1.

By the law of total expectation

$$\mathbb{E}[\mathbb{1}_{\neg \mathcal{M}_1} \hat{N}^q_{X^1,e_j}] \tag{129}$$

$$= \sum^\infty_{t=1} \mathbb{P}(T_1 = t) \mathbb{P}(\neg \mathcal{M}_1 \mid T_1 = t) \mathbb{E}[\hat{N}^q_{X^1,e_j} \mid T_1 = t, \neg \mathcal{M}_1].$$

The union bound implies that $\mathbb{P}(\mathcal{A} \mid \cup_i \mathcal{B}_i) \le \sum_i \mathbb{P}(\mathcal{A} \mid \mathcal{B}_i)$ for any countable collection of events $\mathcal{A}, \mathcal{B}_i$ with $\mathbb{P}(\mathcal{B}_i) \ne 0$. Observe that

$$\neg \mathcal{M}_1 = \cup_{t_0 \in \mathbb{Z}_{\ge 1}} \{E_{X^1,t_0} = e_1 \text{ and } t_0 \le T_1\} \cup \{E_{Z^1,t_0} = e_1 \text{ and } t_0 \le T_1\}. \tag{130}$$

Therefore, since $\hat{N}^q_{X^1,e_j} = \sum_{t_1=1}^{T_1} \mathbb{1}_{E_{X^1,t_1}=e_j}$

$$\mathbb{E}[\hat{N}^q_{X^1,e_j} \mid T_1 = t, \neg\mathcal{M}_1]$$

$$= \sum_{t_1,\ldots,t_q=1}^{t} \mathbb{P}(E_{X^1,t_1} = e_j, \ldots, E_{X^1,t_q} = e_j \mid T_1 = t, \neg\mathcal{M}_1) \tag{131}$$

$$\leq \sum_{t_1,\ldots,t_q=1}^{t} \Big(\sum_{t_0} \mathbb{P}(E_{X^1,t_1} = e_j \mid T_1 = t, E_{X^1,t_0} = e_1) \tag{132}$$

$$+ \sum_{t_0} \mathbb{P}(E_{X^1,t_1} = e_j \mid T_1 = t, E_{Z^1,t_0} = e_1) \Big) \tag{133}$$

where the second sum in (132) runs over all $t_0 \in \{1, \ldots, t\}$ with $\mathbb{P}(T_1 = t, E_{X^1,t_0} = e_1) \neq 0$ and the sum in (133) runs over all $t_0 \in \{1, \ldots, t\}$ with $\mathbb{P}(T_1 = t, E_{Z^1,t_0} = e_1) \neq 0$. First consider the terms with $E_{Z^1,t_0} = e_1$. Note that there are at least $\alpha_{min}^2 n^2$ edges e whose starting point and ending point have the same clusters as the starting point and ending point of e_j respectively. Conditional on $T_1 = t$ and $E_{Z^1,t_0} = e_1$ any such edge e is equally likely to be traversed at time t_1 by X^1. It follows that $\mathbb{P}(E_{X^1,t_1} = e_j \mid T_1 = t, E_{Z^1,t_0} = e_1) \leq \alpha_{min}^{-2} n^{-2}$.

Let us now consider the terms with $E_{X^1,t_0} = e_1$. When $|t_0 - t_1| > 1$ the foregoing argument applies word-for-word and yields that $\mathbb{P}(E_{X^1,t_1} = e_j \mid T_1 = t, E_{X^1,t_0} = e_1) \leq \alpha_{min}^{-2} n^{-2}$. The cases $t_0 = t_1 - 1$ and $t_0 = t_1 + 1$ require a modification.

When $t_0 = t_1 - 1$ note that there at least $\alpha_{min} n$ edges e whose ending point is in the same cluster as the ending point of e_j and whose starting point is equal to the starting point of e_j. Conditional on $T_1 = t$ and $E_{X^1,t_0} = e_1$ any such edge e is equally likely to be traversed at time t_1. Hence, $\mathbb{P}(E_{X^1,t_1} = e_j \mid T_1 = t, E_{X^1,t_0} = e_1) \leq \alpha_{min}^{-1} n^{-1}$ for all $t_0 = t_1 - 1$.

Similarly, when $t_0 = t_1 + 1$ note that there are at least $\alpha_{min} n$ edges e whose starting point is in the same cluster as the starting point of e_j and whose ending point is equal to the ending point of e_j. Conditional on $T_1 = t$ and $E_{X^1,t_0} = e_1$ any such edge e is equally likely to be traversed at time t_1. Hence, $\mathbb{P}(E_{X^1,t_1} = e_j \mid T_1 = t, E_{X^1,t_0} = e_1) \leq \alpha_{min}^{-1} n^{-1}$ for all $t_0 = t_1 + 1$.

Using these bounds in (132) yields that

$$\mathbb{E}[\hat{N}^q_{X^1,e_j} \mid T_1 = t, \neg\mathcal{M}_1] \leq 2\alpha_{min}^{-2} n^{-1} t^{q+1}. \tag{134}$$

Observe that $\mathbb{P}(\neg\mathcal{M}_1 \mid T_1 = t) \leq \mathbb{E}[\hat{N}_{X^1,e_1} + \hat{N}_{Z^1,e_1} \mid T_1 = t]$. Note that X^1 and Z^1 follow the same distribution and apply (124) with $q = 1$ to deduce that

$$\mathbb{P}(\neg\mathcal{M}_1 \mid T_1 = t) \leq 2\mathbb{E}[\hat{N}_{X^1,e_1} \mid T_1 = t] \leq 2\alpha_{min}^{-2} n^{-2} t \tag{135}$$

Hence, (129), (134) and (135) yield that

$$\mathbb{E}[\mathbb{1}_{\neg\mathcal{M}_1} \hat{N}^q_{X^1,e_j}] \leq 4\alpha_{min}^{-4} n^{-3} \mathbb{E}[T_1^{q+2}] \tag{136}$$

where, as in Part 3.1, it holds that $\mathbb{E}[T_1^{q+2}]$ is a finite constant which does not depend on n. Hence, $\mathbb{E}[\mathbb{1}_{\neg\mathcal{M}_1} \hat{N}^q_{X^1,e_j}] = \Theta_q(n^{-3})$ which may be combined with (128) and the fact that $\ell = \Theta(n^{-2})$ to deduce the desired result:

$$\mathbb{E}\left[\left(\sum_{i=1}^{\ell} \mathbb{1}_{\neg\mathcal{M}_i} \hat{N}_{X^i,e_j}\right)^q\right] = O_q(n^{-1}). \tag{137}$$
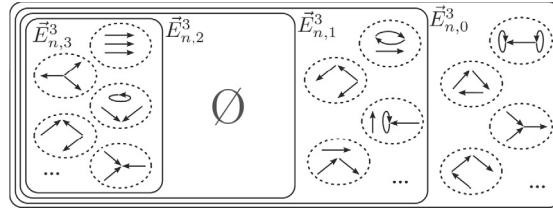
This concludes the proof. $\square$

Fig. 8. Visualization of some configurations of edges occurring in $\vec{E}_{n,r'}^{R}$ for varying values of r' when $R = 3$. The visualized configurations are not exhaustive. Observe that $\vec{E}_{n,R}^{R} = \vec{E}_{n,R-1}^{R} \subseteq \vec{E}_{n,R-2}^{R} \subseteq \cdots \subseteq \vec{E}_{n,0}^{R}$ as is immediate from the definition.

Proposition 6.15. *Assume that X starts in equilibrium and let $0 \le r \le R$ be positive integers. Then, for any $m \in \mathbb{Z}_{\ge 0}^{R}$ with $m_i = 1$ for $i = 1, \ldots, r$, it holds that as n tends to infinity*

$$\max_{\forall i \ne j : e_i \ne e_j} \left| \mathbb{E}\left[M_{X,e_1}^{m_1} \cdots M_{X,e_R}^{m_R} \right] \right| = O_{m,R}(n^{-\lceil r/2 \rceil}) \tag{138}$$

where the maximum runs over all sequences of distinct edges $e_1, \ldots, e_R \in \vec{E}_n$.

We will prove a better bound than Proposition 6.15. Recall from Section 2.4 that [33] improved the convergence in probability of [40] to convergence almost surely provided additional assumptions. It turns out that the most immediate generalization of their results would be too restrictive to include block Markov chains because the correlations in Proposition 6.15 do not decay sufficiently quickly.

The correlation between edges e_i and e_j is maximal when the ending point of e_i is equal to the starting point of e_j. Correspondingly, Proposition 6.15 can be improved when there are many edges whose ending point is not the starting point of some other edge. This improvement could be a point of departure to strengthen Theorems 1.1 and 1.2 to convergence almost surely.

For any positive integers $0 \le r' \le R$ let $\vec{E}_{n,r'}^{R}$ denote the collection of sequences of distinct edges $e_1, \ldots, e_R \in \vec{E}_n$ such that for every $i \in \{1, \ldots, r'\}$ it holds that the ending point of e_i is not the starting point of any edge e_j with $j \in \{1, \ldots, R\} \setminus \{i\}$ and the starting point of e_i is not the ending point of any edge e_j with $j \in \{1, \ldots, R\} \setminus \{i\}$.

See Fig. 8 for an example with $R = 3$. The following proposition includes Proposition 6.15 as the special case with $r' = 0$.

Proposition 6.16. *Assume that X starts in equilibrium and let $0 \le r' \le r \le R$ be positive integers. Then, for any $m \in \mathbb{Z}_{\ge 0}^{R}$ with $m_i = 1$ for $i = 1, \ldots, r$, it holds that as n tends to infinity*

$$\max_{(e_1, \ldots, e_R) \in \vec{E}_{n,r'}^{R}} \left| \mathbb{E}\left[M_{X,e_1}^{m_1} \cdots M_{X,e_R}^{m_R} \right] \right| = O_{m,R}(n^{-\lceil r/2 \rceil - \lceil r'/2 \rceil}). \tag{139}$$

Proof. This proof combines the proof of Proposition 4.8 with an inductive argument as was used in the proof of Proposition 6.14. The main technical difference is that we can no longer use the Cauchy–Schwarz inequality during the inductive step as in (118): the resulting squares would reduce r to zero which weakens the conclusion of the induction hypothesis. Instead, the step analogous to (118) will employ conditional independence. The price we pay for this argument is that it necessitates a stronger induction hypothesis to account for the added conditioning.

The same preliminary reductions as in the proof of Proposition 6.14 are applicable. Firstly, precisely as in Part 1 of the proof of Proposition 4.8, it can be assumed that $K \geq 2R + 1$ by splitting a cluster into pieces of asymptotically equal proportions. Further, by fixing an isomorphism type for the labeled directed graph G induced by $\{e_1, \ldots, e_R\}$ we may again pretend that the edges $e_1, \ldots, e_R$ stay fixed as n tends to infinity.

Part 1: Set-up of the inductive argument

Recall that it is ensured that $K \geq 2R + 1$. Hence, there exists some $k \in \{1, \ldots, K\}$ such that $\mathcal{V}_k$ does not contain any endpoint of e_j for $j = 1, \ldots, R$.

For any $d \in \mathbb{Z}_{\geq 0}$, $\ell' \leq \ell$ and $\tau \in \{0, \ldots, \ell'\}^d$ denote $\mathcal{V}_{X,\tau}$ for the event where $X_{\tau_i} \in \mathcal{V}_k$ for every $i = 1, \ldots, d$. When $d = 0$ it is to be understood that $\mathcal{V}_{X,\tau}$ refers to the universal event. In particular, $\mathbb{P}(\mathcal{V}_{X,\tau}) = 1$ in this case. Fix some d, ℓ' and τ with $\mathbb{P}(\mathcal{V}_{X,\tau}) > 0$ and let $Y := (Y_t)_{t=0}^{\ell'}$ be a sample path from the block Markov chain conditioned on the event $\mathcal{V}_{Y,\tau}$. We will show that there exists a constant $c_{19} \in \mathbb{R}_{\geq 0}$, depending on d but not on ℓ' or τ, such that

$$\mathbb{E}[M_{Y,e_1}^{m_1} \cdots M_{Y,e_R}^{m_R}] \leq c_{19} n^{-\lceil r/2 \rceil - \lceil r'/2 \rceil}. \tag{140}$$

Taking $d = 0$ and $\ell' = \ell$ then recovers the proposition.

The proof of the claim proceeds by induction on r. This is why we require (140) to hold for any $d \in \mathbb{Z}_{\geq 0}$ and $\ell' \leq \ell$ even though the proposition only concerns the case with $d = 0$ and $\ell' = \ell$: we will modify d and ℓ' when reducing r in the inductive step. The argument for the base case $r = 0$ is provided in Part 4.1. Now let $r > 1$ and assume that (140) is known to hold for any smaller value of r.

Recall that $m_1 = 1$ and $\hat{N}_{Y,e_1} = \sum_{t_1=1}^{\ell'} \mathbb{1}_{E_{Y,t_1} = e_1}$. Therefore,

$$\mathbb{E}[M_{Y,e_1}^{m_1} \cdots M_{Y,e_R}^{m_R}] = \mathbb{E}[\hat{N}_{Y,e_1} M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}] - \mathbb{E}[\hat{N}_{Y,e_1}] \mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}]$$

$$= \sum_{t_1=1}^{\ell'} \left(\mathbb{E}[\mathbb{1}_{E_{Y,t_1} = e_1} M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}] - \mathbb{E}[\mathbb{1}_{E_{Y,t_1} = e_1}] \mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}] \right)$$

$$= \sum_{t_1=1}^{\ell'} \mathbb{P}(E_{Y,t_1} = e_1) \left(\mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R} \mid E_{Y,t_1} = e_1] - \mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}] \right). \tag{141}$$

Note that there are at least $\alpha_{min}^2 n^2$ edges e for which the starting and ending point are in the same clusters as the starting and ending point of e_1 respectively. These edges e are equally likely to be traversed at time t_1 by Y. It follows that

$$\mathbb{P}(E_{Y,t_1} = e_1) \leq \alpha_{min}^{-2} n^{-2}. \tag{142}$$

It remains to show that the effect of conditioning on $E_{Y,t_1} = e_1$ in (141) has a small effect on $M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}$. That is, it has to be shown that $\mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R} \mid E_{Y,t_1} = e_1] \approx \mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}]$ with approximation error uniform over all $t_1 \in \{1, \ldots, \ell'\}$ with $\mathbb{P}(E_{Y,t_1} = e_1) > 0$. Fix such a value of t_1.

Part 2: Construction of chains (Y, Y', Z)

Use the following procedure to construct a triple of chains (Y, Y', Z) with Y, Z of length $\ell' + 1$ and Y' of a random length at most $\ell' + 1$. See also Fig. 9 for a visualization of the construction.

(i) Sample an infinitely long path $(\tilde{Y}_t)_{t=-\infty}^{\infty}$ from the block Markov chain conditioned on $\mathcal{V}_{\tilde{Y},\tau}$. Sample $(\tilde{W}_t)_{t=-\infty}^{\infty}$ as an infinitely long path from the block Markov chain
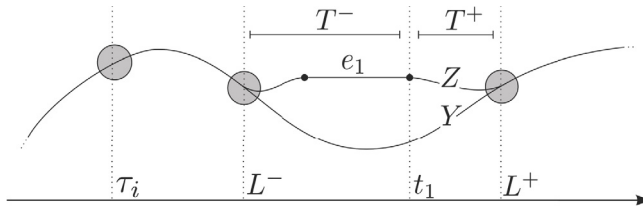
Fig. 9. Visualization of the construction of the chains Y and Z in the proof of Proposition 6.16. The chain Y' is found from Y by cutting out the piece between L^- and L^+. Both chains are conditioned to be in cluster $\mathcal{V}_k$ at times τ_i but Z has the additional condition of using edge e_1 at time t_1. The visualized process of gluing e_1 onto Y exploits the cluster structure to ensure that $L^+ - L^-$ is small.

conditioned on $E_{\widetilde{W},t_1} = e_1$ and $\mathcal{V}_{\widetilde{W},\tau}$. Note that it is possible to sample at negative times by use of a time reversal which exists by the assumption that the Markov chain associated with p is acyclic and irreducible.

(ii) Define

$$T^- := t_1 - \sup\{t \in \mathbb{Z}_{<t_1} : \widetilde{Y}_t \in \mathcal{V}_k, \widetilde{W}_t \in \mathcal{V}_k\} \tag{143}$$

$$T^+ := \inf\{t \in \mathbb{Z}_{>t_1} : \widetilde{Y}_t \in \mathcal{V}_k, \widetilde{W}_t \in \mathcal{V}_k\} - t_1 \tag{144}$$

and note that these values are finite with probability one by the assumption that the Markov chain associated with p is irreducible and acyclic. Let $L^- := \max\{0, t_1 - T^-\}$ and $L^+ := \min\{\ell', t_1 + T^+\}$.

(iii) Let $Y := (\widetilde{Y}_t)_{t=0}^{\ell'}$ and define $Z := (Z_t)_{t=0}^{\ell'}$ by $Z_t = \widetilde{W}_t$ for $t \in \{L^-, \dots, L^+\}$ and $Z_t = \widetilde{Y}_t$ otherwise. Define Y' to be the path of length $\ell' - (L^+ - L^-) + 1 \le \ell + 1$ found by concatenating $(\widetilde{Y}_t)_{t=0}^{L^-}$ and $(\widetilde{Y}_t)_{t=L^++1}^{\ell'}$. Thus, Y' is equal to Y except for the fact that a piece was cut out.

Observe that Y and Z are paths of length $\ell' + 1$ from the block Markov chain conditioned on $\mathcal{V}_{Y,\tau}$ and the block Markov chain conditioned on $\mathcal{V}_{Z,\tau}$ and $E_{Z,t_1} = e_1$, respectively. Correspondingly, $\hat{N}_{Z,e_j}$ has the same distribution as $\hat{N}_{Y,e_j}$ conditioned on the event $E_{Y,t_1} = e_1$.

Part 3: Bounding $\mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R} \mid E_{Y,t_1} = e_1] - \mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}]$

We continue with the notation of the previous part, in particular t_1 and τ are still considered to be fixed.

The distribution of Y' is obscure unless one conditions on L^- and L^+: even the length of Y' is random without this conditioning. In particular, the usual notation $M_X = \hat{N}_X - \mathbb{E}[\hat{N}_X]$ is not so useful for Y' because $\mathbb{E}[\hat{N}_{Y'}]$ averages out the critical dependence on L^+ and L^-. For this reason it is convenient to define a variant of $M_{Y'}$ which takes L^- and L^+ into account:

$$\mathring{M}_{Y'} := \hat{N}_{Y'} - \mathbb{E}[\hat{N}_{Y'} \mid L^+, L']. \tag{145}$$

For $j = 2, \dots, R$ define $\Delta_{Y,j} := \hat{N}_{Y,e_j} - \hat{N}_{Y',e_j}$, $\Delta_{Z,j} := \hat{N}_{Z,e_j} - \hat{N}_{Y',e_j}$ and set

$$\mathcal{E}_{Y,j} := \Delta_{Y,j} - (\mathbb{E}[\hat{N}_{Y,e_j}] - \mathbb{E}[\hat{N}_{Y',e_j} \mid L^-, L^+]), \tag{146}$$

$$\mathcal{E}_{Z,j} := \Delta_{Z,j} - (\mathbb{E}[\hat{N}_{Y,e_j}] - \mathbb{E}[\hat{N}_{Y',e_j} \mid L^-, L^+]). \tag{147}$$

Observe that $M_{Y,e_j} = \mathring{M}_{Y',e_j} + \mathcal{E}_{Y,j}$. Moreover, M_{Y,e_j} conditioned on $E_{Y,t_1} = e_1$ has the same distribution as $\hat{N}_{Z,e_j} - \mathbb{E}[\hat{N}_{Y,e_j}] = \mathring{M}_{Y',e_j} + \mathcal{E}_{Z,j}$. Correspondingly,

$$
\mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R} | E_{Y,t_1} = e_1] - \mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}] \tag{148}
$$
$$
= \mathbb{E}\left[\prod_{j=2}^{R} (\mathring{M}_{Y',e_j} + \mathcal{E}_{Z,j})^{m_j} - \prod_{j=2}^{R} (\mathring{M}_{Y',e_j} + \mathcal{E}_{Y,j})^{m_j} \right].
$$

The key observation is that the leading-order terms in the expansion of the right-hand side of (148) cancel out. This is to say that

$$
\mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R} | E_{Y,t_1} = e_1] - \mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}] \tag{149}
$$
$$
= \sum_{0 \leq m' \leq m} c_{m'} \left(\mathbb{E}\left[\prod_{j=2}^{R} \mathcal{E}_{Z,j}^{m'_j} \mathring{M}_{Y',e_j}^{m_j - m'_j} \right] - \mathbb{E}\left[\prod_{j=2}^{R} \mathcal{E}_{Y,j}^{m'_j} \mathring{M}_{Y',e_j}^{m_j - m'_j} \right] \right)
$$

where the summation runs over vectors of positive integers $m' \in \mathbb{Z}_{\geq 0}^r$ and the $c_{m'}$ are absolute constants with $c_{m'} = 0$ if $m'_j = 0$ for all $j \in \{2, \ldots, R\}$.

Fix some $m' \in \mathbb{Z}^r$ with $m'_j \neq 0$ for some $j \in \{2, \ldots, R\}$. We will consider the terms with $\mathcal{E}_{Z,j}$ in (149); those with $\mathcal{E}_{Y,j}$ may be treated identically. By the law of total expectation,

$$
\mathbb{E}\left[\prod_{j=2}^{R} \mathcal{E}_{Z,j}^{m'_j} \mathring{M}_{Y',e_j}^{m_j - m'_j} \right] = \sum_{\ell^+ > \ell^-} \mathbb{P}\left(\begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix} \right) \mathbb{E}\left[\prod_{j=2}^{R} \mathcal{E}_{Z,j}^{m'_j} \mathring{M}_{Y',e_j}^{m_j - m'_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix} \right]. \tag{150}
$$

The compressed notation for conditional expectation employed in (150) was formally introduced in Section 3.5. Recall that $\mathcal{V}_k$ was chosen not to contain any endpoint of $e_1, \ldots, e_R$. In particular $\Delta_{Z,j}$ is a function of the values of $\widetilde{Z}_t$ for $t \in \{L^- + 1, \ldots, L^+ - 1\}$ whereas $\mathring{M}_{Y',e_j}$ is a function of the values of $\widetilde{Y}_t$ with $t \in \{0, \ldots, L^- - 1, L^+ + 1, \ldots, \ell'\}$. Further, conditional on $L^- = \ell^-$ and $L^+ = \ell^+$ it holds that $\mathbb{E}[\hat{N}_{Y',e_j} | L^-, L^+]$ is a deterministic scalar. Recall the definition for $\mathcal{E}_{Z,j}$ in (147) and conclude that $\prod_{j=2}^{R} \mathcal{E}_{Z,j}^{m'_j}$ is conditionally independent of $\prod_{j=2}^{R} \mathring{M}_{Y',e_j}^{m_j - m'_j}$ given $L^- = \ell^-$ and $L^+ = \ell^+$. Hence,

$$
\mathbb{E}\left[\prod_{j=2}^{R} \mathcal{E}_{Z,j}^{m'_j} \mathring{M}_{Y',e_j}^{m_j - m'_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix} \right] \tag{151}
$$
$$
= \mathbb{E}\left[\prod_{j=2}^{R} \mathcal{E}_{Z,j}^{m'_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix} \right] \mathbb{E}\left[\prod_{j=2}^{R} \mathring{M}_{Y',e_j}^{m_j - m'_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix} \right].
$$

We will use the induction hypothesis (140) to deal with $\prod_{j=2}^{R} \mathring{M}_{Y',e_j}^{m_j - m'_j}$. Let Y'' denote a random path which is distributed as Y' conditioned on $L^- = \ell^-$ and $L^+ = \ell^+$. Recall the notation for $\mathring{M}_{Y',e_j}$ in (145) and conclude that $\prod_{j=2}^{R} \mathring{M}_{Y',e_j}^{m_j - m'_j}$ conditioned on $L^- = \ell^-$ and $L^+ = \ell^+$ has the same distribution as $\prod_{j=2}^{R} M_{Y'',e_j}^{m_j - m'_j}$. Observe that Y'' is a path of length $\ell' - (\ell^+ - \ell^-) + 1$ from the block Markov chain conditioned on $\mathcal{V}_{Y'',\tau'}$ where τ' is the vector of length $d + 1$ such that $\tau'_{d+1} = \ell^-$ and for $i = 1, \ldots, d$ it holds that $\tau'_i = \tau_i$ if $\tau_i \in \{0, \ldots, \ell^-\}$ and $\tau'_i = \max\{\ell^- + \tau_i - \ell^+, \ell^-\}$ otherwise. The induction hypothesis (140) is applicable and, by

inspection of the exponents $m_i - m'_i$ with $i \in \{1, \ldots, r'\}$, we may conclude that

$$\mathbb{E}\left[\prod_{j=2}^{R} \mathring{M}_{Y',e_j}^{m_j-m'_j} \;\middle|\; \begin{array}{l} L^- = \ell^- \\ L^+ = \ell^+ \end{array}\right] = O_{m,R,d}\left(n^{-f_1(r,m')-f_2(r,r',m')}\right) \tag{152}$$

with $f_1(r, m') := \lceil (r - \#\{j \in \{2, \ldots, r\} : m'_j = 1\} - 1)/2 \rceil$ and $f_2(r, r', m') := \lceil (r' - \#\{j \in \{2, \ldots, r'\} : m'_j = 1\} - 1)/2 \rceil$.

In Part 4.3 we will show that there exists some constant $c_{20} \in \mathbb{R}_{>0}$, which does not depend on t_1 or τ, such that

$$\left|\mathbb{E}\left[\prod_{j=2}^{R} \mathcal{E}_{Z,j}^{m'_j} \;\middle|\; \begin{array}{l} L^- = \ell^- \\ L^+ = \ell^+ \end{array}\right]\right| \leq c_{20} n^{-g_1(m')-g_2(m',r')}(\ell^+ - \ell^-)^{3\|m'\|_1} \tag{153}$$

where $g_1(m') := \#\{j \in \{2, \ldots, R\} : m'_j \neq 0\}$ and

$$g_2(m', r') = \begin{cases} 0 & \text{if } g_1(m') = 0 \text{ or } r' = 0, \\ g_3(m', r') + 1 & \text{if } 0 \leq g_3(m', r') < g_1(m') \text{ and } r' > 0, \\ g_3(m', r') & \text{if } 0 < g_3(m', r') = g_1(m') \text{ and } r' > 0, \end{cases} \tag{154}$$

where $g_3(m', r') := \#\{j \in \{2, \ldots, r'\} : m'_j \neq 0\}$. It will then follow from (150)–(153) that there exists a constant $c_{21} \in \mathbb{R}_{>0}$ such that

$$\left|\mathbb{E}\left[\prod_{j=2}^{R} \mathcal{E}_{Z,j}^{m'_j} \mathring{M}_{Y',e_j}^{m_j-m'_j}\right]\right| \tag{155}$$

$$\leq c_{21} n^{-(f_1(r,m')+g_1(m'))-(f_2(r,r',m')+g_2(m',r'))} \mathbb{E}[(L^+ - L^-)^{3\|m'\|_1}].$$

Observe that for any positive integers $q_1, q_2 \in \mathbb{Z}_{\geq 0}$ it holds that $\lceil (q_1 - q_2 - 1)/2 \rceil + \max\{1, q_2\} \geq \lceil q_1/2 \rceil$. The condition below (149) stating that $c_{m'} = 0$ when $m'_j = 0$ for all $j \in \{2, \ldots, R\}$ ensures that $g_1(m') \geq \max\{1, \#\{j \in \{2, \ldots, r\} : m'_j = 1\}\}$ for all terms with $c_{m'} \neq 0$. Therefore, using the definition of f_1,

$$f_1(r, m') + g_1(m') \geq \lceil r/2 \rceil \tag{156}$$

for all terms with $c_{m'} \neq 0$ in (149). Note that $f_2(r, r', m') + g_2(m', r') = 0$ when $r' = 0$. If $r' > 0$, note that $g_2(m', r') \geq \max\{1, \#\{j \in \{2, \ldots, r'\} : m'_j = 1\}\}$ for all terms in (149) with $c_{m'} \neq 0$. Hence, using the definition of f_2,

$$f_2(r, r', m') + g_2(m', r') \geq \lceil r'/2 \rceil \tag{157}$$

for all terms with $c_{m'} \neq 0$ in (149).

It will be shown in Part 4.2 that $\mathbb{E}[(T^+)^q] = O_{q,d}(1)$ for any $q \in \mathbb{Z}_{\geq 0}$ and a similar conclusion holds for T^-. Note that it is here also claimed that the bound is uniform in t_1 and τ provided that d is fixed. Now observe that $(L^+ - L^-)^{3\|m'\|_1} \leq (T^+ + T^-)^{3\|m'\|_1}$. Expand $(T^+ + T^-)^{3\|m'\|_1}$ and apply the Cauchy–Schwarz inequality to the resulting monomial terms to derive that $\mathbb{E}[(L^+ - L^-)^{3\|m'\|_1}] = O_{m',d}(1)$.

It now follows by (149) and (155)–(157) that there exists a constant $c_{22} \in \mathbb{R}_{>0}$ such that

$$\mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R} \mid E_{Y,t_1} = e_1] - \mathbb{E}[M_{Y,e_2}^{m_2} \cdots M_{Y,e_R}^{m_R}] \leq c_{22} n^{-\lceil r/2 \rceil - \lceil r'/2 \rceil}. \tag{158}$$

Given that there are $\ell' \leq \ell = \Theta(n^2)$ terms in (141) and that $\mathbb{P}(E_{Y,e_1} = e_1) = O(n^{-2})$ by (142) it follows that there exists a constant $c_{23} \in \mathbb{R}_{>0}$ such that

$$|\mathbb{E}[M_{Y,e_1}^{m_1} \cdots M_{Y,e_R}^{m_R}]| \leq c_{23} n^{-\lceil r/2 \rceil - \lceil r'/2 \rceil} \tag{159}$$

which is the desired result.

Part 4: Remaining Bounds

Part 4.1: Base case: $r = 0$

Recall that we still have to prove that (140) holds for the base case $r = 0$. By repeated application of the Cauchy–Schwarz inequality the claim is reduced to the statement that $\mathbb{E}[M_{Y,e_j}^q] = O_{q,d}(1)$ for any $j \in \{2, \ldots, R\}$ and $q \in \mathbb{Z}_{\geq 0}$. By definition $M_{Y,e_j} = \hat{N}_{Y,e_j} - \mathbb{E}[\hat{N}_{Y,e_j}]$ so it suffices to show that $\mathbb{E}[\hat{N}_{Y,e_j}^{q'}] = O_{q',d}(1)$ for any $q' \in \mathbb{Z}_{\geq 0}$. Further, since $\hat{N}_{Y,e_j}$ can only increase when one extends Y to a longer path it may be assumed that $\ell' = \ell$. In this case Y follows the same distribution as X conditioned on the event $\mathcal{V}_{X,\tau}$.

We claim that for any fixed d there exists a constant $c_{24} \in \mathbb{R}_{>0}$ independent of n such that for any $\tau \in \{0, \ldots, \ell'\}^d$ with $\mathbb{P}(\mathcal{V}_{X,\tau}) > 0$ it holds that

$$\mathbb{P}(\mathcal{V}_{X,\tau}) \geq c_{24}. \tag{160}$$

Indeed, assume without loss of generality that the τ_i are nondecreasing in i. Since X starts in equilibrium it then holds that

$$\mathbb{P}(\mathcal{V}_{X,\tau}) = \mathbb{P}(X_{\tau_1} \in \mathcal{V}_k) \prod_{i=1}^{d-1} \mathbb{P}(X_{\tau_{i+1}} \in \mathcal{V}_k \mid X_{\tau_i \in \mathcal{V}_k}) \tag{161}$$

$$= \pi(k) \prod_{i=1}^{d-1} p^{\tau_{i+1} - \tau_i}(k, k). \tag{162}$$

This implies (160) since, by the Markov chain associated with p being irreducible and acyclic, it holds that $p^t(k, k)$ tends to $\pi(k)$ as t tends to infinity and it holds that $\pi(k) > 0$.

By the law of total expectation it follows that

$$\mathbb{E}[\hat{N}_{Y,e_j}^{q'}] = \mathbb{E}[\hat{N}_{X,e_j}^{q'} \mid \mathcal{V}_{X,\tau}] \tag{163}$$

$$\leq c_{24}^{-1} \mathbb{E}[\hat{N}_{X,e_j}^{q'}] \tag{164}$$

which establishes the desired result since $\mathbb{E}[\hat{N}_{X,e_j}^{q'}] = O_{q'}(1)$ by Corollary 6.13.

Part 4.2: $\mathbb{E}[(T^+)^q] = O_{q,d}(1)$

Let $q \in \mathbb{Z}_{\geq 0}$ and recall the definition of T^+ in (144). Consider the product chain $\Sigma_{(\tilde{Y}, \tilde{W})}^+ = (\sigma(\tilde{Y}_{t_1+t}), \sigma(\tilde{W}_{t_1+t}))_{t=0}^\infty$ on the space of clusters $\{1, \ldots, K\} \times \{1, \ldots, K\}$. Then T^+ is the first strictly positive time $\Sigma_{(\tilde{Y}, \tilde{W})}^+$ is in (k, k). Sample infinitely long sample paths $V := (V_t)_{t=0}^\infty$ and $V' := (V_t')_{t=0}^\infty$ from the block Markov chain with V' conditioned on V_{t_1}' being the ending point of e_1. Consider the product chain $\Sigma_{(V,V')}^+ = (\sigma(V_{t_1+t}), \sigma(V_{t_1+t}'))_{t=0}^\infty$ on the space of clusters $\{1, \ldots, K\} \times \{1, \ldots, K\}$. Denote $T_{V,V'}^+$ for the first strictly positive time $\Sigma_{(V,V')}^+$ is in (k, k). By the Markov chain associated with p being irreducible and acyclic it holds that $\mathbb{P}(T_{V,V'}^+ > t)$ shows exponential decay in t. In particular, $\mathbb{E}[(T_{V,V'}^+)^q]$ is finite and independent of n.

Now observe that $\Sigma_{(\tilde{Y}, \tilde{W})}^+$ has the same distribution as $\Sigma_{(V,V')}^+$ conditioned on the events $\mathcal{V}_{V,\tau}$ and $\mathcal{V}_{V',\tau}$. By (160) there exists a constant $c_{24} \in \mathbb{R}_{>0}$ independent of n and t_1 such that $\mathbb{P}(\mathcal{V}_{V,\tau}) \geq c_{24}$ for all $\tau \in \{0, \ldots, \ell'\}^d$ with $\mathbb{P}(\mathcal{V}_{V,\tau}) > 0$. It can similarly be deduced that there exists a constant $c_{25} \in \mathbb{R}_{>0}$ such that $\mathbb{P}(\mathcal{V}_{V',\tau}) \geq c_{25}$ for all $\tau \in \{0, \ldots, \ell'\}^d$ with $\mathbb{P}(\mathcal{V}_{V',\tau}) > 0$.

Now, by the law of total expectation and the fact that $\mathcal{V}_{V',\tau}$ is independent of $\mathcal{V}_{V,\tau}$

$$\mathbb{E}[(T^+)^q] = \mathbb{E}[(T^+_{V,V'})^q \mid \mathcal{V}_{V,\tau}, \mathcal{V}_{V',\tau}] \tag{165}$$

$$\leq (c_{24}c_{25})^{-1}\mathbb{E}[(T^+_{V,V'})^q] \tag{166}$$

which establishes the desired result.

Part 4.3: $|\mathbb{E}[\prod_{j=2}^R \mathcal{E}_{Z,j}^{q_j} \mid L^- = \ell^-, L^+ = \ell^+]| \leq c_{20}n^{-g_1(q)-g_2(q,r')}(\ell^+ - \ell^-)^{3\|q\|_1}$

Consider some fixed $q \in \mathbb{Z}_{\geq 0}^R$ and $\ell^-, \ell^+ \in \mathbb{Z}_{\geq 0}$ with $\mathbb{P}(L^- = \ell^-, L^+ = \ell^+) > 0$. Let us remark that the following arguments also apply when Z is replaced by Y. In fact, the process Y is easier to deal with since Z is conditioned to be at e_1 at time t_1.

Recall the definition for $\mathcal{E}_{Z,j}$ in (147). Note that conditional on $L^+ = \ell^+$ and $L^- = \ell^-$, it holds that $\mathbb{E}[\hat{N}_{Y,e_j}] - \mathbb{E}[\hat{N}_{Y',e_j} \mid L^-, L^+]$ is a deterministic scalar. Therefore, after substituting the definition of $\mathcal{E}_{Z,j}$ in $\prod_{j=2}^R \mathcal{E}_{Z,j}^{q_j}$ and by then expanding in terms of $\Delta_{Z,j}$ and $\mathbb{E}[\hat{N}_{Y,e_j}] - \mathbb{E}[\hat{N}_{Y',e_j} \mid L^-, L^+]$ it follows that

$$\mathbb{E}\left[\prod_{j=2}^R \mathcal{E}_{Z,j}^{q_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right] \tag{167}$$

$$= \sum_{0 \leq q' \leq q} c_{q'} \mathbb{E}\left[\prod_{j=2}^R \Delta_{Z,j}^{q_j'} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right] \prod_{j=2}^R \left(\mathbb{E}\left[\hat{N}_{Y,e_j}\right] - \mathbb{E}\left[\hat{N}_{Y',e_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right]\right)^{q_j - q_j'}$$

for certain absolute constants $c_{q'}$. The sum runs here over vectors of integers of length R.

Fix some $0 \leq q' \leq q$ and let us establish a bound on $\mathbb{E}[\prod_{j=2}^R \Delta_{Z,j}^{q_j'} \mid L^- = \ell^-, L^+ = \ell^+]$. Observe that the product $\prod_{j=2}^R \Delta_{Z,j}^{q_j'}$ can only be nonzero if $\Delta_{Z,j} \neq 0$ for every j with $q_j' \neq 0$. Recall from (154) that $g_1(q')$ is the number of nonzero q_j' with $j \in \{2, \ldots, R\}$ and $g_3(q', r')$ is the number of nonzero q_j' with $j \in \{2, \ldots, r'\}$. Hence the product $\prod_{j=2}^R \Delta_{Z,j}^{q_j'}$ can only be nonzero if $\sum_{j=2}^R \mathbb{1}_{\Delta_{Z,j}>0} \geq g_1(q')$ and $\sum_{j=2}^{r'} \mathbb{1}_{\Delta_{Z,j}>0} \geq g_3(q', r')$. Therefore, by the law of total expectation

$$\mathbb{E}\left[\prod_{j=2}^R \Delta_{Z,j}^{q_j'} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right] = \mathbb{P}\left(\begin{matrix} \sum_{j=2}^R \mathbb{1}_{\Delta_{Z,j}>0} \geq g_1(q') \\ \sum_{j=2}^{r'} \mathbb{1}_{\Delta_{Z,j}>0} \geq g_3(q', r') \end{matrix} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right) \tag{168}$$

$$\times \mathbb{E}\left[\prod_{j=2}^R \Delta_{Z,j}^{q_j'} \,\middle|\, \begin{matrix} L^- = \ell^- & \sum_{j=2}^R \mathbb{1}_{\Delta_{Z,j}>0} \geq g_1(q') \\ L^+ = \ell^+, & \sum_{j=2}^{r'} \mathbb{1}_{\Delta_{Z,j}>0} \geq g_3(q', r') \end{matrix}\right].$$

It follows from $\prod_{j=2}^R \Delta_{Z,j}^{q_j'} \leq (L^+ - L^-)^{\|q'\|_1}$ that

$$\mathbb{E}\left[\prod_{j=2}^R \Delta_{Z,j}^{q_j'} \,\middle|\, \begin{matrix} L^- = \ell^- & \sum_{j=2}^R \mathbb{1}_{\Delta_{Z,j}>0} \geq g_1(q') \\ L^+ = \ell^+, & \sum_{j=2}^{r'} \mathbb{1}_{\Delta_{Z,j}>0} \geq g_3(q', r') \end{matrix}\right] \leq (\ell^+ - \ell^-)^{\|q'\|_1}. \tag{169}$$

The goal thus becomes to establish a bound on the probability on the right-hand side of (168). This bound will be established in two steps. First, we show that the event described by the probability implies that $(Z_t)_{t \in \{\ell^-, \ldots, \ell^+\} \setminus \{t_1, t_1 - 1\}}$ has to visit the endpoints of $e_2, \ldots, e_R$ often. This step is achieved in (182). Second, we show that visiting many endpoints is a rare event. This step is achieved in (190).
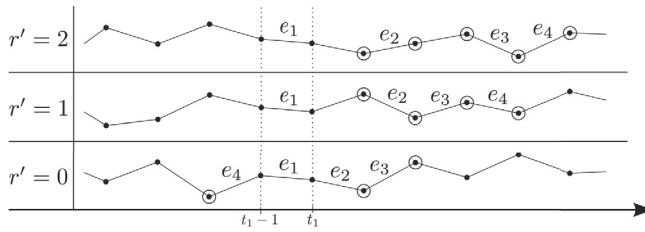
Fig. 10. Visualization of some 'worst case' realizations of Z for which $\sum_{j=2}^{R} \mathbb{1}_{\mathcal{T}_{e_j} \neq \emptyset} \geq g_1(q')$ and $\sum_{j=2}^{r'} \mathbb{1}_{\mathcal{T}_{e_j} \neq \emptyset} \geq g_3(q', r')$ when $g_1(q') = R - 1 = 3$ and $g_3(q', r') = r'$ varies. The points contributing to $\#(\cup_{j=2}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\})$ are circled. Note that the number of circled points increases as r' does so. Indeed, $r' > 0$ avoids losing points due to the exclusion of $\{t_1, t_1 - 1\}$ whereas $r' \geq j$ with $j \geq 2$ avoids losses due to the possibility that e_j shares endpoints with the other e_i.

For every $j \in \{1, \ldots, R\}$ consider the following collections of times which measure when we are in e_j or its endpoints

$$\mathcal{T}_{e_j} := \{t \in \{\ell^- + 1, \ldots, \ell^+\} : E_{Z,t} = e_j\}, \qquad \mathcal{T}'_{e_j} := \cup_{t \in \mathcal{T}_{e_j}} \{t, t - 1\}. \tag{170}$$

Observe that in this notation it holds that $\Delta_{Z,j} = \#\mathcal{T}_{e_j}$. Hence, the probability on the right-hand side of (168) may be rewritten as

$$\mathbb{P}\left(\begin{array}{c} \sum_{j=2}^{R} \mathbb{1}_{\Delta_{Z,j} > 0} \geq g_1(q') \\ \sum_{j=2}^{r'} \mathbb{1}_{\Delta_{Z,j} > 0} \geq g_3(q', r') \end{array} \middle| \begin{array}{c} L^- = \ell^- \\ L^+ = \ell^+ \end{array} \right) \tag{171}$$

$$= \mathbb{P}\left(\begin{array}{c} \sum_{j=2}^{R} \mathbb{1}_{\mathcal{T}_{e_j} \neq \emptyset} \geq g_1(q') \\ \sum_{j=2}^{r'} \mathbb{1}_{\mathcal{T}_{e_j} \neq \emptyset} \geq g_3(q', r') \end{array} \middle| \begin{array}{c} L^- = \ell^- \\ L^+ = \ell^+ \end{array} \right).$$

We claim that whenever the event described in the probability on the right-hand side of (171) holds, it follows that $\#(\cup_{j=2}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\}) \geq g_1(q') + g_2(q', r')$. The key difficulty is that the sets $\mathcal{T}'_{e_1}, \ldots, \mathcal{T}'_{e_R}$ may not be disjoint due to the fact that the e_j can share endpoints; see Fig. 10.

Recall the definition of $\vec{E}_{n,r'}^R$ from the paragraph preceding Proposition 6.16. It follows that for every $(i, j) \in \{1, \ldots, R\} \times \{1, \ldots, r'\}$ with $i \neq j$ it holds that $\mathcal{T}'_{e_i} \cap \mathcal{T}'_{e_j} = \emptyset$. Hence,

$$\#\left(\bigcup_{j=2}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\} \right) \tag{172}$$

$$= \sum_{j=2}^{r'} \#\left(\mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\} \right) + \#\left(\bigcup_{j=r'+1}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\} \right) \tag{173}$$

$$= \sum_{j=2}^{r'} \#\mathcal{T}'_{e_j} + \#\left(\bigcup_{j=r'+1}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\} \right) \tag{174}$$

where in (174) we used that $\{t_1, t_1 - 1\} \subseteq \mathcal{T}'_{e_1}$.

Note that for every $j \in \{2, \ldots, R\}$ with $\mathcal{T}_{e_j} \neq \emptyset$ it holds that $\#\mathcal{T}'_{e_j} \geq 2$. Thus, it follows that

$$\sum_{j=2}^{r'} \#\mathcal{T}'_{e_j} \geq 2 \sum_{j=2}^{r'} \mathbb{1}_{\mathcal{T}_{e_j} \neq \emptyset}. \tag{175}$$

When $r' > 0$ it holds that $\mathcal{T}'_{e_1}$ is disjoint from $\mathcal{T}'_{e_j}$ for every $j \in \{2, \ldots, R\}$. In particular $\{t_1, t_1 - 1\}$ is disjoint from $\mathcal{T}'_{e_j}$ for every $j \in \{2, \ldots, R\}$. Hence,

$$\bigcup_{j=r'+1}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\} = \bigcup_{j=r'+1}^{R} \mathcal{T}'_{e_j}, \quad \text{if } r' > 0. \tag{176}$$

Whenever $\cup_{j=r'+1}^{R} \mathcal{T}_{e_j} \neq \emptyset$ we can construct a subset of $\cup_{j=r'+1}^{R} \mathcal{T}'_{e_j}$ as

$$\left\{ \min\left\{ t - 1 : t \in \bigcup_{j=r'+1}^{R} \mathcal{T}_{e_j} \right\} \right\} \cup \bigcup_{j=r'+1}^{R} \mathcal{T}_{e_j} \subseteq \bigcup_{j=r'+1}^{R} \mathcal{T}'_{e_j}. \tag{177}$$

Observe that the left-hand side of (177) is a union of disjoint sets due to the fact that $e_2, \ldots, e_R$ are distinct edges. It now follows from (176) that if $r' > 0$,

$$\#\left(\bigcup_{j=r'+1}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\} \right) \geq \mathbb{1}_{(\cup_{j=r'+1}^{R} \mathcal{T}_{e_j}) \neq \emptyset} + \sum_{j=r'+1}^{R} \#\mathcal{T}_{e_j} \tag{178}$$

$$\geq \mathbb{1}_{(\cup_{j=r'+1}^{R} \mathcal{T}_{e_j}) \neq \emptyset} + \sum_{j=r'+1}^{R} \mathbb{1}_{\mathcal{T}_{e_j} \neq \emptyset}. \tag{179}$$

When $r' = 0$ we still have an injection from $\cup_{j=r'+1}^{R} \mathcal{T}_{e_j}$ into $\cup_{j=r'+1}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\}$ defined by $t \mapsto t\mathbb{1}_{t > t_1} + (t - 1)\mathbb{1}_{t < t_1}$. Hence,

$$\#\left(\bigcup_{j=r'+1}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\} \right) \geq \#\left(\bigcup_{j=r'+1}^{R} \mathcal{T}_{e_j} \right) \geq \sum_{j=r'+1}^{R} \mathbb{1}_{\mathcal{T}_{e_j} \neq \emptyset}. \tag{180}$$

Combine (174)–(180) to deduce that

$$\#\left(\bigcup_{j=2}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\} \right) \tag{181}$$

$$\geq \begin{cases} \sum_{j=2}^{R} \mathbb{1}_{\mathcal{T}_{e_j} \neq \emptyset} + \sum_{j=2}^{r'} \mathbb{1}_{\mathcal{T}_{e_j} \neq \emptyset} + \mathbb{1}_{(\cup_{j=r'+1}^{R} \mathcal{T}_{e_j}) \neq \emptyset} & \text{if } r' > 0, \\ \sum_{j=2}^{R} \mathbb{1}_{\mathcal{T}_{e_j} \neq \emptyset} & \text{if } r' = 0. \end{cases}$$

Recall the condition described in the probability on the right-hand side of (171). Note that whenever this condition holds it follows that $\sum_{j=2}^{R} \mathbb{1}_{\mathcal{T}_{e_j} \neq \emptyset} \geq g_1(q')$ and $\sum_{j=2}^{r'} \mathbb{1}_{\mathcal{T}_{e_j} \neq \emptyset} + \mathbb{1}_{(\cup_{j=r'+1}^{R} \mathcal{T}_{e_j}) \neq \emptyset} \geq g_3(q', r') + \mathbb{1}_{g_1(q') > g_3(q', r')}$. Further, recall from (154) that the definition of g_2 in terms of g_3 has three cases. Now, using (181) in (171) and checking each case from the

definition of g_2 individually we find that

$$\mathbb{P}\left(\begin{matrix} \sum_{j=2}^{R} \mathbb{1}_{\Delta_{Z,j}>0} \geq g_1(q') \\ \sum_{j=2}^{r'} \mathbb{1}_{\Delta_{Z,j}>0} \geq g_3(q',r') \end{matrix} \;\middle|\; \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right) \tag{182}$$
$$\leq \mathbb{P}\left(\#\left(\bigcup_{j=2}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\}\right) \geq g_1(q') + g_2(q',r') \;\middle|\; \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right).$$

Let us denote $V(e_2, \ldots, e_R)$ for the set of cardinality $\leq 2(R-1)$ consisting of the endpoints of the edges $e_2, \ldots, e_R$. Observe that $\#(\cup_{j=2}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\}) \geq g_1(q') + g_2(q',r')$ implies that there exists some $T' \subseteq \{\ell^-, \ldots, \ell^+\} \setminus \{t_1, t_1 - 1\}$ with $\#T' = g_1(q') + g_2(q',r')$ such that $Z_{t'} \in V(e_2, \ldots, e_R)$ for all $t' \in T'$. Consequently, by the union bound

$$\mathbb{P}\left(\#\left(\bigcup_{j=2}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\}\right) \geq g_1(q') + g_2(q',r') \;\middle|\; \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right) \tag{183}$$
$$\leq \sum_{\#T'=g_1(q')+g_2(q',r')} \mathbb{P}\left(\{Z_{t'} : t' \in T'\} \subseteq V(e_2, \ldots, e_R) \;\middle|\; \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right) \tag{184}$$

where the sum runs over all subsets $T' \subseteq \{\ell^-, \ldots, \ell^+\} \setminus \{t_1, t_1 - 1\}$ with $\#T' = g_1(q') + g_2(q',r')$. Fix such a subset T' and denote $\Sigma_Z := (\sigma(Z_t))_{t=0}^{\ell}$ for chain of clusters associated to Z. Then, the law of total probability yields that

$$\mathbb{P}\left(\{Z_{t'} : t' \in T'\} \subseteq V(e_2, \ldots, e_R) \;\middle|\; \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right) \tag{185}$$
$$= \sum_{s_Z} \mathbb{P}\left(\Sigma_Z = s_Z \;\middle|\; \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right) \mathbb{P}\left(\{Z_{t'} : t' \in T'\} \subseteq V(e_2, \ldots, e_R) \;\middle|\; \Sigma_Z = s_Z\right)$$

where the sum runs over all $s_Z \subseteq \{1, \ldots, K\}^{\ell'+1}$ with $\mathbb{P}(\Sigma_Z = s_Z \mid L^- = \ell^-, L^+ = \ell^+) \neq 0$. Now, observe that $(Z_t)_{t \in \{0, \ldots, \ell'\} \setminus \{t_1, t_1 - 1\}}$ is uniformly distributed in $\prod_{t \in \{0, \ldots, \ell'\} \setminus \{t_1, t_1 - 1\}} \mathcal{V}_{s_{Z,t}}$ given that $\Sigma_Z = s_Z$. Hence, by using that $\#V(e_2, \ldots, e_R) \leq 2(R-1)$ and $\#\mathcal{V}_k \geq \alpha_{min} n$ for every $k \in \{1, \ldots, K\}$

$$\mathbb{P}(\{Z_{t'} : t' \in T'\} \subseteq V(e_2, \ldots, e_R) \mid \Sigma_Z = s_Z) \tag{186}$$
$$= \prod_{t' \in T'} \frac{\#(V(e_2, \ldots, e_R) \cap \mathcal{V}_{s_{Z,t'}})}{\#\mathcal{V}_{s_{Z,t'}}} \tag{187}$$
$$\leq (2(R-1)\alpha_{min}^{-1} n^{-1})^{g_1(q')+g_2(q',r')} \tag{188}$$
$$\leq c_{26} n^{-g_1(q')-g_2(q',r')} \tag{189}$$

where we defined $c_{26} := \max_{0 \leq q' \leq q} (2(R-1)\alpha_{min}^{-1})^{g_1(q')+g_2(q',r')}$. Now, by (183)–(189) with the bound $\binom{\ell^+ - \ell^- - 1}{g_1(q')+g_2(q',r')} \leq (\ell^+ - \ell^-)^{g_1(q')+g_2(q',r')}$ for the number of terms in the sum of (183)

$$\mathbb{P}\left(\#\left(\bigcup_{j=2}^{R} \mathcal{T}'_{e_j} \setminus \{t_1, t_1 - 1\}\right) \geq g_1(q') + g_2(q',r') \;\middle|\; \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right) \tag{190}$$
$$\leq c_{26} n^{-(g_1(q')+g_2(q',r'))} (\ell^+ - \ell^-)^{g_1(q')+g_2(q',r')}.$$

Combine (168), (169) and (182) with the bound (190) to conclude that

$$\mathbb{E}\left[\prod_{j=2}^{R} \Delta_{Z,j}^{q_j'} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right] \leq c_{26}(\ell^+ - \ell^-)^{3\|q'\|_1} n^{-(g_1(q')+g_2(q',r'))} \tag{191}$$

where it was used that $g_1(q') \leq \|q'\|_1$ and $g_2(q') \leq \|q'\|_1$. This establishes the desired upper bound for the conditional expectation of $\prod_{j=2}^{R} \Delta_{Z,j}^{q_j}$.

Let us remark that the only property about the chain Z which was used in the foregoing argument is that $(Z_t)_{t \in \{0,\dots,\ell'\}\setminus\{t_1,t_1-1\}}$ is uniformly distributed in $\prod_{t \in \{0,\dots,\ell'\}\setminus\{t_1,t_1-1\}} \mathcal{V}_{s_Z}$ when conditioned on $\Sigma_Z = s_Z$. Hence, the conclusion (191) also applies to other chains with this property such as Y. This is to say that by repeating the argument for (191) word-for-word one finds a constant $c_{27} \in \mathbb{R}_{>0}$ such that

$$\mathbb{E}\left[\prod_{j=2}^{R} \Delta_{Y,j}^{q_j'} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right] \leq c_{27}(\ell^+ - \ell^-)^{3\|q'\|_1} n^{-(g_1(q')+g_2(q',r'))} \tag{192}$$

for any $0 \leq q' \leq q$.

Next, we consider the factors $(\mathbb{E}[\hat{N}_{Y,e_j}] - \mathbb{E}[\hat{N}_{Y',e_j} \mid L^- = \ell^-, L^+ = \ell^+])$ in (167). Fix some $j \in \{2,\dots,R\}$ and observe that

$$\mathbb{E}[\hat{N}_{Y,e_j}] - \mathbb{E}\left[\hat{N}_{Y',e_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right] = \left(\mathbb{E}[\hat{N}_{Y,e_j}] - \mathbb{E}\left[\hat{N}_{Y,e_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right]\right)$$
$$+ \left(\mathbb{E}\left[\hat{N}_{Y,e_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right] - \mathbb{E}\left[\hat{N}_{Y',e_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right]\right). \tag{193}$$

Here, since $\hat{N}_{Y,e_j} - \hat{N}_{Y',e_j} = \Delta_{Y,j}$, the bound (192) yields a constant $c_{28} \in \mathbb{R}_{>0}$ such that

$$\mathbb{E}\left[\hat{N}_{Y,e_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right] - \mathbb{E}\left[\hat{N}_{Y',e_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right] = \mathbb{E}\left[\Delta_{Y,j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right]. \tag{194}$$

$$\leq \begin{cases} c_{28}n^{-1}(\ell^+ - \ell^-)^3 & \text{if } r' = 0, \\ c_{28}n^{-2}(\ell^+ - \ell^-)^3 & \text{if } r' > 0. \end{cases} \tag{195}$$

It remains to consider the term $\mathbb{E}[\hat{N}_{Y,e_j}] - \mathbb{E}[\hat{N}_{Y,e_j} \mid L^- = \ell^-, L^+ = \ell^+]$ in (193). This term may be studied by means of a coupling argument. Construct a path G of length $\ell' + 1$ by using the following procedure:

(i) Let $\widetilde{Y} := (\widetilde{Y}_t)_{t=-\infty}^{\infty}$ be the path used in the construction of Y in Part 2. Independently sample $\widetilde{G} := (\widetilde{G}_t)_{t=-\infty}^{\infty}$ from the same distribution as $\widetilde{Y}$ conditioned on $L^- = \ell^-$ and $L^+ = \ell^+$.

(ii) Define

$$\widetilde{T}^- = \ell^- - \sup\{t \in \mathbb{Z}_{<\ell^-} : \widetilde{Y}_t \in \mathcal{V}_k, \widetilde{G} \in \mathcal{V}_k\} \tag{196}$$
$$\widetilde{T}^+ = \inf\{t \in \mathbb{Z}_{>\ell^+} : \widetilde{Y}_t \in \mathcal{V}_k, \widetilde{G}_t \in \mathcal{V}_k\} - \ell^+ \tag{197}$$

and note that these values are finite with probability one by the assumption that the Markov chain associated with p is irreducible and acyclic. Let $\widetilde{L}^- = \max\{0, \ell^- - \widetilde{T}^-\}$ and $\widetilde{L}^+ = \min\{\ell', \ell^+ + \widetilde{T}^+\}$.

(iii) Define $G := (G_t)_{t=0}^{\ell'}$ by $G_t := \widetilde{G}_t$ for $t \in \{\widetilde{L}^-, \dots, \widetilde{L}^+\}$ and $G_t = \widetilde{Y}_t$ otherwise.

Note that G is here implicitly dependent on ℓ^- and ℓ^+ but this is suppressed in the notation. Indeed, G has the same distribution as Y conditioned on $L^- = \ell^-$ and $L^+ = \ell^+$. For any $j \in \{2, \ldots, R\}$ let $\widetilde{\Delta}_{G,j}$ and $\widetilde{\Delta}_{Y,j}$ denote the number of times $(G_t)_{t=\widetilde{L}^-}^{\widetilde{L}^+}$ and $(Y_t)_{t=\widetilde{L}^-}^{\widetilde{L}^+}$ traversed edge e_j. Then,

$$\left| \mathbb{E}\left[\hat{N}_{Y,e_j}\right] - \mathbb{E}\left[\hat{N}_{Y,e_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right]\right| = \left|\mathbb{E}\left[\hat{N}_{Y,e_j} - \hat{N}_{G,e_j}\right]\right| \tag{198}$$

$$\leq |\mathbb{E}[\widetilde{\Delta}_{Y,j}]| + |\mathbb{E}[\widetilde{\Delta}_{G,j}]|. \tag{199}$$

As in (192) it may be established that for any $\widetilde{\ell}^+, \widetilde{\ell}^- \in \mathbb{Z}_{\geq 0}$ with $\mathbb{P}(\widetilde{L}^+ = \widetilde{\ell}^+, \widetilde{L}^- = \widetilde{\ell}^-) > 0$ it holds that

$$\left|\mathbb{E}\left[\widetilde{\Delta}_{Y,j} \,\middle|\, \begin{matrix} \widetilde{L}^- = \widetilde{\ell}^- \\ \widetilde{L}^+ = \widetilde{\ell}^+ \end{matrix}\right]\right| \leq \begin{cases} c_{28} n^{-1}(\widetilde{\ell}^+ - \widetilde{\ell}^-)^3 & \text{if } r' = 0, \\ c_{28} n^{-2}(\widetilde{\ell}^+ - \widetilde{\ell}^-)^3 & \text{if } r' > 0. \end{cases} \tag{200}$$

By the law of total expectation it now follows that

$$|\mathbb{E}[\widetilde{\Delta}_{Y,j}]| \leq \begin{cases} c_{28} n^{-1}\mathbb{E}[(\widetilde{L}^+ - \widetilde{L}^-)^3] & \text{if } r' = 0, \\ c_{28} n^{-2}\mathbb{E}[(\widetilde{L}^+ - \widetilde{L}^-)^3] & \text{if } r' > 0, \end{cases} \tag{201}$$

and a similar conclusion applies to $\widetilde{\Delta}_{G,j}$.

We claim that $\mathbb{E}[(\widetilde{T}^+)^m] = O_m(1)$ and similarly $\mathbb{E}[(\widetilde{T}^-)^m] = O_m(1)$ for any $m \in \mathbb{Z}_{\geq 0}$. Indeed, this may be deduced as in (166) by consideration of $\Sigma^+_{(\widetilde{Y},\widetilde{G})} := (\sigma(\widetilde{Y}_{\ell^++t}), \sigma(\widetilde{G}_{\ell^++t}))_{t=0}^\infty$, $\widetilde{V} := (\widetilde{V}_t)_{t=0}^\infty$ and $\widetilde{V}' := (\widetilde{V}_t)_{t=0}^\infty$ where $\widetilde{V}$ is sample path from the block Markov chain and $\widetilde{V}'$ is sample path from the block Markov chain conditioned on the event $\widetilde{V}_{\ell^+} \in \mathcal{V}_k$. Next, use a binomial expansion on $\widetilde{L}^+ - \widetilde{L}^- = \ell^+ - \ell^- + (\widetilde{T}^+ + \widetilde{T}^-)$ to write

$$\mathbb{E}[(\widetilde{L}^+ - \widetilde{L}^-)^3] = \sum_{m=0}^{3} \binom{3}{m} \mathbb{E}[(\widetilde{T}^+ + \widetilde{T}^-)^m](\ell^+ - \ell^-)^{3-m} \tag{202}$$

$$= \sum_{m=0}^{3} \binom{3}{m} \sum_{m'=0}^{m} \binom{m}{m'} \mathbb{E}[(\widetilde{T}^+)^{m'}(\widetilde{T}^-)^{m-m'}](\ell^+ - \ell^-)^{3-m} \tag{203}$$

The coefficients $\mathbb{E}[(\widetilde{T}^+)^{m'}(\widetilde{T}^-)^{m-m'}]$ in (203) are $O_{m,m'}(1)$ by the Cauchy–Schwarz inequality. Therefore, since $(\ell^+ - \ell^-) \geq 1$, there exists some constant $c_{29} \in \mathbb{R}_{>0}$ such that

$$\mathbb{E}[(\widetilde{L}^+ - \widetilde{L}^-)^3] \leq c_{29}(\ell^+ - \ell^-)^3. \tag{204}$$

By (199)—(204) it follows that there exists some constant $c_{30} \in \mathbb{R}_{>0}$ such that

$$\left| \mathbb{E}\left[\hat{N}_{Y,e_j}\right] - \mathbb{E}\left[\hat{N}_{Y,e_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right]\right| \leq \begin{cases} c_{30} n^{-1}(\ell^+ - \ell^-)^3 & \text{if } r' = 0, \\ c_{30} n^{-2}(\ell^+ - \ell^-)^3 & \text{if } r' > 0. \end{cases} \tag{205}$$

By (193), (195) and (205) we may conclude that

$$\left| \mathbb{E}[\hat{N}_{Y,e_j}] - \mathbb{E}\left[\hat{N}_{Y',e_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right]\right| \leq \begin{cases} c_{31} n^{-1}(\ell^+ - \ell^-)^3 & \text{if } r' = 0, \\ c_{31} n^{-2}(\ell^+ - \ell^-)^3 & \text{if } r' > 0. \end{cases} \tag{206}$$

for some constant $c_{31} \in \mathbb{R}_{>0}$. Combine (191) and (206) to find a constant $c_{32} \in \mathbb{R}_{>0}$ such that for any $0 \leq q' \leq q$

$$\mathbb{E}\left[\prod_{j=2}^{R} \Delta_{Z,j}^{q'_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right] \prod_{j=2}^{R}\left(\mathbb{E}\left[\hat{N}_{Y,e_j}\right] - \mathbb{E}\left[\hat{N}_{Y',e_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right]\right)^{q_j - q'_j} \tag{207}$$

$$\leq \begin{cases} c_{32} n^{-g_1(q') - g_2(q',r') - \sum_{j=2}^{R}(q_j - q'_j)}(\ell^+ - \ell^-)^{3\|q\|_1} & \text{if } r' = 0, \\ c_{32} n^{-g_1(q') - g_2(q',r') - 2\sum_{j=2}^{R}(q_j - q'_j)}(\ell^+ - \ell^-)^{3\|q\|_1} & \text{if } r' > 0, \end{cases} \tag{208}$$

where it was used that $\|q'\|_1 + \|q - q'\|_1 = \|q\|_1$. Recall the definition of g_1 and g_2 in (154) and observe that each additional nonzero coordinate in q' increases $g_1(q') + g_2(q',r')$ by at most $1 + \mathbb{1}_{r'>0}$. Hence, the worst bound in (208) is attained at $q' = q$ and we may conclude that

$$\mathbb{E}\left[\prod_{j=2}^{R} \Delta_{Z,j}^{q'_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right] \prod_{j=2}^{R}\left(\mathbb{E}\left[\hat{N}_{Y,e_j}\right] - \mathbb{E}\left[\hat{N}_{Y',e_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right]\right)^{q_j - q'_j} \tag{209}$$

$$\leq c_{32} n^{-g_1(q) - g_2(q,r')}(\ell^+ - \ell^-)^{3\|q\|_1}. \tag{210}$$

Plug (210) into (167) to find a constant $c_{33} \in \mathbb{R}_{>0}$ such that

$$\left| \mathbb{E}\left[\prod_{j=2}^{R} \mathcal{E}_{Z,j}^{q_j} \,\middle|\, \begin{matrix} L^- = \ell^- \\ L^+ = \ell^+ \end{matrix}\right]\right| \leq c_{33} n^{-g_1(q) - g_2(q,r')}(\ell^+ - \ell^-)^{3\|q\|_1} \tag{211}$$

which is the desired result. □

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E. Abbe, Community detection and stochastic block models: recent developments, J. Mach. Learn. Res. 18 (1) (2017) 6446–6531.

[2] K. Adhikari, I. Jana, K. Saha, Linear eigenvalue statistics of random matrices with a variance profile, Random Matrices Theory Appl. 10 (03) (2021) 2250004.

[3] O. Ajanki, L. Erdős, T. Krüger, Quadratic Vector Equations on Complex Upper Half-Plane, No. 1261, American Mathematical Society, 2019.

[4] J. Aljadeff, D. Renfrew, M. Stern, Eigenvalues of block structured asymmetric random matrices, J. Math. Phys. 56 (10) (2015) 103502.

[5] G.W. Anderson, A. Guionnet, O. Zeitouni, An Introduction to Random Matrices, Cambridge University Press, 2010.

[6] R. Arratia, L. Goldstein, L. Gordon, Two moments suffice for Poisson approximations: the Chen–Stein method, Ann. Probab. (1989) 9–25.

[7]   A. Athreya, D.E. Fishkind, M. Tang, C.E. Priebe, Y. Park, J.T. Vogelstein, K. Levin, V. Lyzinski, Y. Qin, Statistical inference on random dot product graphs: a survey, J. Mach. Learn. Res. 18 (1) (2017) 8393–8484.

[8]   Z. Bai, J.W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, Vol. 20, Springer, 2010.

[9]   M. Banna, Limiting spectral distribution of Gram matrices associated with functionals of β-mixing processes, J. Math. Anal. Appl. 433 (1) (2016) 416–433.

[10] M. Banna, F. Merlevède, Limiting spectral distribution of large sample covariance matrices associated with a class of stationary processes, J. Theoret. Probab. 28 (2) (2015) 745–783.

[11] M. Banna, F. Merlevède, P. Youssef, Bernstein-type inequality for a class of dependent random matrices, Random Matrices Theory Appl. 5 (02) (2016) 1650006.

[12] M. Bhattacharjee, A. Bose, Large sample behaviour of high-dimensional autocovariance matrices, Ann. Statist. 44 (2) (2016) 598–628.

[13] C. Bordenave, P. Caputo, D. Chafaï, Spectrum of large random reversible Markov chains: two examples, ALEA Lat. Am. J. Probab. Math. Stat. 7 (2010) 41–64.

[14] C. Bordenave, P. Caputo, D. Chafaï, Spectrum of large random reversible Markov chains: heavy-tailed weights on the complete graph, Ann. Probab. 39 (4) (2011) 1544–1590.

[15] C. Bordenave, P. Caputo, D. Chafaï, Circular law theorem for random Markov matrices, Probab. Theory Related Fields 152 (3) (2012) 751–779.

[16] C. Bordenave, P. Caputo, D. Chafaï, D. Piras, Spectrum of large random Markov chains: heavy-tailed weights on the oriented complete graph, Random Matrices Theory Appl. 6 (02) (2017) 1750006.

[17] A. Bose, W. Hachem, Smallest singular value and limit eigenvalue distribution of a class of non-Hermitian random matrices with statistical application, J. Multivariate Anal. 178 (2020) 104623.

[18] A. Bose, W. Hachem, Spectral measures of empirical autocovariance matrices of high-dimensional Gaussian stationary processes, 2021, arXiv preprint arXiv:2110.08523.

[19] R. Catalano, M. Fleermann, W. Kirsch, Random band and block matrices with correlated entries, 2022, arXiv preprint arXiv:2202.04707.

[20] D. Chafaï, Aspects of large random Markov kernels, Stochastics 81 (3–4) (2009) 415–429.

[21] D. Chafaï, The Dirichlet Markov ensemble, J. Multivariate Anal. 101 (3) (2010) 555–567.

[22] A. Chakrabarty, R.S. Hazra, D. Sarkar, From random matrices to long range dependence, Random Matrices Theory Appl. 5 (02) (2016) 1650008.

[23] A. Chatterjee, R.S. Hazra, Spectral properties for the Laplacian of a generalized Wigner matrix, Random Matrices Theory Appl. (2021) 2250026.

[24] L.H. Chen, Poisson approximation for dependent trials, Ann. Probab. 3 (3) (1975) 534–545.

[25] Z. Chi, Random reversible Markov matrices with tunable extremal eigenvalues, Ann. Appl. Probab. 26 (4) (2016) 2257–2272.

[26] N. Cook, W. Hachem, J. Najim, D. Renfrew, Non-Hermitian random matrices with a variance profile (I): deterministic equivalents and limiting ESDs, Electron. J. Probab. 23 (2018) 1–61.

[27] X. Ding, Spectral analysis of large block random matrices with rectangular blocks, Lith. Math. J. 54 (2) (2014) 115–126.

[28] X. Ding, On some spectral properties of large block Laplacian random matrices, Statist. Probab. Lett. 99 (2015) 61–69.

[29] W. Doeblin, Exposé de la théorie des chaînes simples constantes de Markov à un nombre fini d'états, Revue Math. Union Interbalkanique 2 (77–105) (1938) 78–80.

[30] Z. Du, N. Ozay, L. Balzano, Mode clustering for Markov jump systems, in: IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, CAMSAP, IEEE, 2019, pp. 126–130.

[31] Y. Duan, T. Ke, M. Wang, State aggregation learning from Markov transition data, Adv. Neural Inf. Process. Syst. 32 (2019) 4486–4495.

[32] R.R. Far, T. Oraby, W. Bryc, R. Speicher, On slow-fading MIMO systems with nonseparable correlation, IEEE Trans. Inform. Theory 54 (2) (2008) 544–553.

[33] M. Fleermann, W. Kirsch, T. Kriecherbauer, The almost sure semicircle law for random band matrices with dependent entries, Stochastic Process. Appl. 131 (2021) 172–200.

[34] O. Friesen, M. Löwe, The semicircle law for matrices with independent diagonals, J. Theoret. Probab. 26 (4) (2013) 1084–1096.

[35] C. Gao, Z. Ma, A.Y. Zhang, H.H. Zhou, Achieving optimal misclassification proportion in stochastic block models, J. Mach. Learn. Res. 18 (1) (2017) 1980–2024.

[36] G. Grimmett, D. Stirzaker, Probability and Random Processes, third ed., Oxford University Press, 2001.

[37] W. Hachem, P. Loubaton, J. Najim, The empirical distribution of the eigenvalues of a Gram matrix with a given variance profile, Ann. Inst. Henri Poincare B 42 (6) (2006) 649–670.

[38] N. Haydn, Entry and return times distribution, Dyn. Syst. 28 (3) (2013) 333–353.

[39] J.W. Helton, R.R. Far, R. Speicher, Operator-valued semicircular elements: Solving a quadratic matrix equation with positivity constraints, Int. Math. Res. Not. 2007 (9) (2007) rnm086, http://dx.doi.org/10.1093/imrn/rnm086.

[40] W. Hochstättler, W. Kirsch, S. Warzel, Semicircle law for a matrix ensemble with dependent entries, J. Theoret. Probab. 29 (3) (2016) 1047–1068.

[41] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, 1991, http://dx.doi.org/10.1017/CBO9780511840371.

[42] M. Horvat, The ensemble of random Markov matrices, J. Stat. Mech. Theory Exp. 2009 (07) (2009) P07005.

[43] G. Innocentini, M. Novaes, Time-inhomogeneous random Markov chains, J. Stat. Mech. Theory Exp. 2018 (10) (2018) 103202.

[44] B. Jin, C. Wang, B. Miao, M.L. Huang, Limiting spectral distribution of large-dimensional sample covariance matrices generated by VARMA, J. Multivariate Anal. 100 (9) (2009) 2112–2125.

[45] O. Kallenberg, Foundations of Modern Probability, third ed., Springer, 2021.

[46] Z.T. Ke, Y. Ma, X. Lin, Estimation of the number of spiked eigenvalues in a covariance matrix by bulk eigenvalue matching analysis, J. Amer. Statist. Assoc. (2021) 1–19.

[47] R. Kühn, Random matrix spectra and relaxation in complex networks, Acta Phys. Polon. B 46 (2015) 1653–1681.

[48] Z. Li, C. Lam, J. Yao, Q. Yao, On testing for high-dimensional white noise, Ann. Statist. 47 (6) (2019) 3382–3412.

[49] Z. Li, G. Pan, J. Yao, On singular value distribution of large-dimensional autocovariance matrices, J. Multivariate Anal. 137 (2015) 119–140.

[50] H. Liu, A. Aue, D. Paul, On the Marčenko–Pastur law for linear time series, Ann. Statist. 43 (2) (2015) 675–712.

[51] L. Lovász, Large Networks and Graph Limits, Vol. 60, American Mathematical Soc., 2012.

[52] L. Lovász, B. Szegedy, Limits of dense graph sequences, J. Combin. Theory Ser. B 96 (6) (2006) 933–957.

[53] M. Löwe, The semicircle law for matrices with ergodic entries, Statist. Probab. Lett. 141 (2018) 90–95.

[54] M. Löwe, K. Schubert, On the limiting spectral density of random matrices filled with stochastic processes, Random Oper. Stoch. Equ. 27 (2) (2019) 89–105.

[55] F. Merlevède, M. Peligrad, On the empirical spectral distribution for matrices with long memory and independent rows, Stochastic Process. Appl. 126 (9) (2016) 2734–2760.

[56] F. Mosam, D. Vidaurre, E. De Giuli, Breakdown of random matrix universality in Markov models, Phys. Rev. E 104 (2) (2021) 024305.

[57] NYC Taxi and Limousine Commission, TLC trip record data, 2016, https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. (Accessed April 2022).

[58] L.H. Oliveira, M. Novaes, Random stochastic matrices from classical compact Lie groups and symmetric spaces, J. Math. Phys. 60 (12) (2019) 123508.

[59] D. Paulin, Concentration inequalities for Markov chains by Marton couplings and spectral methods, Electron. J. Probab. 20 (2015) 1–32.

[60] C. Peligrad, M. Peligrad, The limiting spectral distribution in terms of spectral density, Random Matrices Theory Appl. 5 (01) (2016) 1650003.

[61] O. Pfaffel, E. Schlemm, Eigenvalue distribution of large sample covariance matrices of linear processes, Probab. Math. Statist. 31 (2012).

[62] O. Pfaffel, E. Schlemm, Limiting spectral distribution of a new random matrix model with dependence across rows and columns, Linear Algebra Appl. 436 (9) (2012) 2966–2979.

[63] B. Pitskel, Poisson limit law for Markov chains, Ergodic Theory Dynam. Systems 11 (3) (1991) 501–513.

[64] A. Rai, P. Shinde, S. Jalan, Network spectra for drug-target identification in complex diseases: new guns against old foes, Appl. Netw. Sci. 3 (1) (2018) 1–18.

[65] W. Rudin, Real and Complex Analysis, third ed., McGraw–Hill Publishing Co., New York, 1987.

[66] J. Sanders, A. Proutière, S. Yun, Clustering in block Markov chains, Ann. Statist. 48 (6) (2020) 3488–3512.

[67] J. Sanders, A. Senen–Cerda, Spectral norm bounds for block Markov chain random matrices, 2021, arXiv preprint arXiv:2111.06201.

[68] C. Sarkar, S. Jalan, Spectral properties of complex networks, Chaos 28 (10) (2018) 102101.

[69] S. Schbath, An overview on the distribution of word counts in Markov chains, J. Comput. Biol. 7 (1–2) (2000) 193–201.

[70] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.

[71] D. Shlyakhtenko, Random Gaussian band matrices and freeness with amalgamation, Int. Math. Res. Not. 1996 (20) (1996) 1013–1025.

[72] K. Stratos, A.M. Rush, S.B. Cohen, M. Collins, Spectral learning of refinement HMMs, in: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, 2013, pp. 56–64.

[73] T. Tao, Topics in random matrix theory, Grad. Stud. Math. 132 (2011).

[74] J. Veraart, D.S. Novikov, D. Christiaens, B. Ades–Aron, J. Sijbers, E. Fieremans, Denoising of diffusion MRI using random matrix theory, Neuroimage 142 (2016) 394–406.

[75] J. Yao, A note on a Marčenko–Pastur type theorem for time series, Statist. Probab. Lett. 82 (1) (2012) 22–28.

[76] J. Yao, W. Yuan, On eigenvalue distributions of large auto-covariance matrices, 2020, arXiv preprint arXiv:2011.09165.

[77] A. Zhang, M. Wang, Spectral state compression of Markov processes, IEEE Trans. Inf. Theory 66 (5) (2020) 3202–3231.

[78] Y. Zhou, A.R. Zhang, L. Zheng, Y. Wang, Optimal high-order tensor SVD via tensor-train orthogonal iteration, IEEE Trans. Inf. Theory (2022).

[79] Y. Zhu, A graphon approach to limiting spectral distributions of Wigner-type matrices, Random Structures Algorithms 56 (1) (2020) 251–279.

[80] Z. Zhu, X. Li, M. Wang, A. Zhang, Learning Markov models via low-rank optimization, Oper. Res. (2021).

[81] K. Zyczkowski, M. Kus, W. Słomczyński, H. Sommers, Random unistochastic matrices, J. Phys. A: Math. Gen. 36 (12) (2003) 3425.