Probabilistic graph inspections through forests

Koperberg, V.T.

Probabilistic Graph Inspections Through Forests

Twan Koperberg

Probabilistic Graph Inspections Through Forests

Twan Koperberg

This dissertation and the research presented therein were supported by the Dutch Research Councel (NWO) through the Gravitation-grant NETWORKS-024.002.003.

Cover: Theresa Song Loong

Probabilistic Graph Inspections Through Forests

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof. dr. S. de Rijcke,
volgens besluit van het college voor promoties
te verdedigen op donderdag 25 juni 2026
klokke 11.30 uur

door

Twan Koperberg
geboren te Amsterdam
in 1989

Promotor:

Prof.dr. W.Th.F. den Hollander

Co-promotores:

Dr. L. Avena (Università degli Studi di Firenze)
Dr. A. Gaudillière (Université d'Aix-Marseille)

Promotiecommissie:

Prof.dr.ir. G.L.A. Derks
Prof.dr. P. Stevenhagen
Prof.dr. N. Enriquez (Université de Paris)
Dr. W. Ruszel (Universiteit Utrecht)
Prof.dr. C. Sabot (Université de Lyon)

Contents

CHAPTER 1
Introduction

Methods for the analysis, interpretation and explanation of the features of large complex networks are both of theoretical importance, e.g. to aid in understanding complex graph models in statistical physics, as well as of applied utility, e.g. since many real-world systems can be modelled as complex networks. Improvements in these methods can provide a better understanding of large network systems. Generally, the difficulty of analyzing a network increases with its complexity.

Forests, defined as acyclic graphs consisting of one or more trees, are elementary networks that exhibit minimal structural complexity. Due to their inherently simple nature, forests are amenable to many forms of analysis. Hence, if the relevant structural properties of a complex network can, at least partially, be captured by a spanning forest of the network, or a collection of spanning forests, then these forests can be a valuable tool in their analysis.

This dissertation consists of two parts, each of which describes a distinct implementation of a complexity reduction scheme with the help of spanning forests.

The topic of part I, which is the largest of the two parts and contains chapters 2 to 4, is a specific probability measure on the rooted spanning forests of an arbitrary given network that depends on a single positive intensity parameter. This measure will be referred to as the *Kirchhoff forest measure*.

The Kirchhoff forest measure is of theoretical relevance due to its connection to various models from statistical physics. It is a variation on the familiar uniform spanning tree, and is further connected to e.g. the random walk loop-soup and the Gaussian free field. Rather than inquiring into thermodynamic limits, the primary focus will be on studying Kirchhoff forests on finite graphs, and, in particular, on the effect of varying the intensity parameter.

While this work focuses on the theoretical study of Kirchhoff forests and does not aim to develop practical applications, possible applications do provide a motivation for this study. One example of a well-studied applied problem in network science is that of community detection, i.e. partitioning the vertices of a network in a manner that emphasizes the connectivity structure of the network. A spanning forest of a network provides a natural way to partition the vertex set, by grouping vertices according to constituent trees. In [7] a random forest-based vertex partitioning was introduced that is based on the Kirchhoff forest measure. Further applications include the estimation of the Laplacian spectrum of a network [9].

The Kirchhoff forest discussed in part I is defined on an arbitrary network. In contrast, the forest reduction exhibited in chapter 5, the sole chapter of part II, is restricted in scope to bipartite graphs that result from a specific problem in coupling theory. As will be shown, for this problem the entire relevant structure of the bipartite graph can be captured by a single deterministic spanning forest.

The remainder of this introductory chapter provides an overview of the relevant mathematical objects for parts I and II in section 1.1 and section 1.2, respectively. In section 1.3 a more detailed description is given of each of the four chapters and of the results presented therein. The final section, section 1.4, concludes this chapter with a summary of several remaining open problems.

1.1 Kirchhoff forests

Given an arbitrary weighted directed graph $\mathcal{G} = (\mathcal{X}, \mathcal{E}, w)$, with vertex set $\mathcal{X}$, directed edge set $\mathcal{E} \subseteq \mathcal{X} \times \mathcal{X}$ and edge-weight function w, the *Kirchhoff forest measure* on $\mathcal{G}$ is a specific measure on the set of rooted forests. More precisely, it is a family of such measures depending on a positive parameter denoted by q. In this work all forests considered will be spanning forests. A *rooted forest* is a forest in which for each of its connected components, i.e. for each tree of the forest, a vertex in that tree is designated as the *root* of that tree. It is notationally convenient to represent a rooted forest as a set of directed edges, with each edge directed towards the root of its tree. In what follows a rooted forest will sometimes be referred to as a graph instead of a collection of edges.

Definition 1.1.1 (Kirchhoff forest). Given an arbitrary edge-weighted directed graph $\mathcal{G} = (\mathcal{X}, \mathcal{E}, w)$, and a parameter $q > 0$, a *Kirchhoff forest* on $\mathcal{G}$ of intensity q is a random rooted forest Φ_q with law

$$\mathbb{P}(\Phi_q = F) := \frac{1}{Z(q)}\, q^{r(F)} \prod_{e \in F} w(e), \quad \text{for all rooted forests } F \text{ of } \mathcal{G},$$

where $r(F)$ denotes the number of roots of F, and $Z(q) := \sum_F q^{r(F)} \prod_{e \in F} w(e)$ is the normalizing partition function. ■

The intensity parameter q controls the expected number of trees of the Kirchhoff forest. For large q a Kirchhoff forest consists of many small trees. For small q a Kirchhoff forest is a rooted forest with few components. The *uniform spanning tree* (UST) measure is recovered in the limit as $q \downarrow 0$. At $q = 1$ the (weighted) uniform measure on rooted forest is obtained.

Over the past decade, Kirchhoff forests have gained increasing attention, both through efforts aimed at finding practical applications, and enlarging theoretical perspectives, see e.g. [1, 10, 20, 70].

1.1.1 Uniform spanning tree

The Kirchhoff forest measure is intimately connected to the UST measure, and hence to various other models related thereto, e.g. the Fortuin-Kasteleyn random-cluster model.

The connection with the UST can be illustrated by considering an extension of the underlying graph on which the Kirchhoff forest measure is defined. This extended

graph is obtained by adding to the original graph a 'cemetery' vertex and directed edges from each vertex towards the cemetery vertex that all have edge-weight q. If T is a (weighted) UST of the extended graph, then the induced subgraph of T obtained by removing the cemetery vertex is a Kirchhoff forest of intensity q on the original graph. The link between the UST and Kirchhoff forest models allows for many results on the UST to be directly adapted to Kirchhoff forests.

The Kirchhoff forest measure should not be confused with another interesting forest measure that is related to the random-cluster model, namely the *uniform spanning forest* mentioned e.g. in [31], which is a measure on the unrooted forests of the graph that has no direct connections to Kirchhoff forests. The uniform spanning forest measure should, in turn, not be mistaken for the UST measure on the integer lattice in dimensions $d > 4$, which, confusingly, is also referred to as the uniform spanning forest, due to the fact that in high dimensions the UST measure is supported on forests [14].

1.1.2 Wilson's algorithm

The primary reason why Kirchhoff forests lend themselves well towards various applied problems is the existence of an efficient sampling procedure due to Wilson [77]. This celebrated procedure, known as *Wilson's algorithm*, allows for the sampling of a Kirchhoff forest Φ_q with expected time complexity $\mathcal{O}(n(1+\frac{\bar{w}}{q}))$, where n is the number of vertices of the graph, and $\bar{w}$ is its mean weighted out-degree [9, 61, 77]. Remarkably, for dense graphs this means that sampling Kirchhoff forests can be done faster than observing all the edges of the network. Recently, Wilson's algorithm has been generalized to a unified framework of efficient sampling procedures known as *partial rejection sampling* [32].

Moreover, Wilson's algorithm is a powerful tool for the theoretical analysis of Kirchhoff forests. It was noted by Pemantle that the law of the path from a given vertex to its root is that of a *loop-erased random walk* [67]. Wilson's procedure makes use of loop-erased random walks to construct a Kirchhoff forest, so that questions on the abstract Kirchhoff forest measure can be translated into questions on random walks, which are often more tangible.

Wilson's algorithm to produce a Kirchhoff forest of intensity q can be described as follows:

- On the graph $\mathcal{G}$ we define a sub-Markovian random walk that is obtained from the random walk on the extended graph, with additional edges of weight q as defined in section 1.1.1 above, by killing the random walk whenever it hits the cemetery vertex.

- Pick an arbitrary vertex and run the sub-Markovian random walk starting from that vertex until it is killed. Each time the random walk makes a cycle, all edges in the cycle are removed from the trajectory of the random walk, to obtain its loop-erased trajectory. This loop-erased trajectory becomes a branch of the Kirchhoff forest.

- Repeatedly pick a new vertex that has not been picked before, and run an independent copy of the sub-Markovian random walk until either it is killed or it hits a vertex in the loop-erased trajectories of any of the previous random walks. Continue in this manner until all vertices have been picked.
- The union of all directed edges in the loop-erased trajectories of the sub-Markovian random walks forms a Kirchhoff forest of intensity q. This intensity is given by the killing rate of the sub-Markovian random walk.

The loop-erased random walks employed in Wilson's procedure have been extensively studied in their own right. See e.g. [53, 54, 74]. For our purposes, a noteworthy result is a theorem due to Marchal [61], that explicitly characterizes the law of a loop-erased random walk as a ratio of determinants, as this result allows for the explicit computation of the probability of obtaining a particular branch in a Kirchhoff forest.

1.1.3 Laplacian spectrum and determinantal processes

An important role in the study of Kirchhoff forests is set aside for the *graph Laplacian matrix* L, i.e. the matrix $(L(x,y))_{x,y\in\mathcal{X}}$ with off-diagonal entries $L(x,y) := -w(x,y)$ and diagonal entries $L(x,x) := \sum_{y\neq x} w(x,y)$. Many quantities related to the Kirchhoff forest measure can be expressed in terms of the Laplacian spectrum. These spectral expressions can be useful in two opposite directions. From a theoretical point of view these expressions can aid computations in cases where the Laplacian spectrum is well understood. Conversely, the spectral expressions can be applied in spectral estimation procedures, relying on the efficient sampling of Kirchhoff forests due to Wilson's algorithm.

The primary example of the relevance of the Laplacian spectrum is Kirchhoff's *matrix-tree theorem*, the eponym of the forest measure. This theorem states that the Kirchhoff forest partition function is given by the characteristic polynomial of $-L$, i.e.

$$Z(q) = \det[qI + L],$$

where I denotes the identity matrix. Another example is the *root process*, i.e. the random subset of vertices that are roots of a Kirchhoff forest Φ_q. The set of roots forms a determinantal process with kernel $K_q := q(qI + L)^{-1}$ [6]. The spectrum of this kernel is easily expressed in terms of the Laplacian spectrum. Moreover, the entries of the matrix K_q have a probabilistic interpretation in terms of Kirchhoff forests [19]: the entry $K_q(x,y)$ is equal to the probability of the event that both x and y belong to the same tree of Φ_q, and y is a root.

Not only the roots of a Kirchhoff forest form a determinantal process. More importantly, when viewed as a subset of the directed edge set a Kirchhoff forest is itself a determinantal process, a result due to Burton and Pemantle known as the *transfer-current theorem* [17].

While a general sampling method exists for determinantal processes with self-adjoint kernels [35], for the case of Kirchhoff forests this method is outperformed by

Wilson's algorithm, which is both more efficient and more generally applicable, not only when the underlying graph $\mathcal{G}$ is undirected, but even in cases where the transfer-current kernel is not self-adjoint. The matrix-tree theorem and the transfer-current theorem ensure that the partition function and the edge correlations can be explicitly computed, and place the Kirchhoff forest model firmly in the realm of integrable systems.

1.1.4 Loop-soups, occupation fields and Gaussian free field

The loops that are removed from the random walk trajectories during Wilson's procedure constitute a random configuration of cycles of the graph. Hence, in addition to producing a Kirchhoff forest Φ_q, Wilson's procedure also generates a random set of cycles $\mathfrak{C}_q$. Remarkably, even though these are obtained from the same random walks, the produced cycles are independent of the resulting Kirchhoff forest [77].

It was shown by Le Jan that the cycle configuration $\mathfrak{C}_q$ produced by Wilson's algorithm is closely related to the *random walk loop-soup* [55], which is a Poisson point process on the countable set of closed walks of the graph $\mathcal{G}$, with intensity measure

$$\mu((x_0, x_1, \dots, x_l)) := \frac{1}{l} \prod_{k=1}^{l} P(x_{k-1}, x_k), \quad \text{for walks } (x_0, x_1, \dots, x_l) \text{ with } x_l = x_0,$$

where P denotes the substochastic transition matrix of the employed sub-Markovian random walk. The link between Wilson's algorithm and the loop-soup is in particular apparent in their respective *occupation fields*. The occupation field $\widehat{C}$ of a (deterministic) set of closed walks C is the map that counts for each vertex x the total number of visits[1] of x by all the closed walks in the set, i.e.

$$\widehat{C}(x) := \sum_{(x_0,x_1,\dots,x_l)\in C} \sum_{k=1}^{l} \mathbf{1}\{x_k = x\} \quad \text{for } x \in \mathcal{X}.$$

This defines the random occupation fields of both $\mathfrak{C}_q$ and the loop-soup. The occupation field of the loop-soup has the same distribution as the occupation field of Wilson's cycle configuration [55].

Le Jan established a further connection between these two occupation fields and the *discrete Gaussian free field* (DGFF) with mass q, which is a centered multi-variate Gaussian random variable $\phi_q = (\phi_q(x))_{x\in\mathcal{X}}$ with covariance matrix given by the Green's function, i.e.

$$\mathbb{E}[\phi_q(x)\phi_q(y)] := (qI + L)^{-1}(x, y), \text{ for } x, y \in \mathcal{X}.$$

For the continuous-time analogue of Wilson's occupation field, Le Jan showed that it is equal in distribution to the average of the squares of two i.i.d. DGFFs ϕ_q and $\tilde{\phi}_q$ with mass q. This connection was further extended by Lupu, who showed how to construct an elegant coupling of the DGFF and the loop-soup [59].

[1] More precisely, the number of arrivals to x is counted, as the starting point x_0 of each closed walk does not contribute to the count.

1.2 Couplings and matchings

The topic of chapter 5 will be two distinct problems, the first related to *couplings* of probability measures, the second concerning *matchings* of bipartite graphs, which are more closely related than might be apparent at first glance.

The following problem regarding couplings of probability measures is considered. Given are two finite sets A and B, two probability measures $\mathbb{P}$ and $\mathbb{P}'$ on A and B, respectively, and a subset $R \subseteq A \times B$ of the product space.

Does there exist a coupling of $\mathbb{P}$ and $\mathbb{P}'$ that is supported on R?

A solution to this problem is provided by Strassen's theorem, which gives a necessary and sufficient condition for the existence of the sought coupling [75]. In the most popular form of Strassen's theorem, the relation R is restricted to being a partial ordering, in which case it is known as *Strassen's theorem on stochastic domination* [57]. The scope of Strassen's thoerem extends beyond this setting, and holds for any closed subset R between two Polish spaces.

In our work, the setting will be restricted to finite sets, in which Strassen's theorem states the following.

Theorem 1.1 (Strassen's theorem for finite sets). *Let A and B be finite sets and $R \subseteq A \times B$ a relation between them. Let $\mathbb{P}$ and $\mathbb{P}'$ be probability measures on A and B, respectively. Then there exists a coupling $\hat{\mathbb{P}}$ of $\mathbb{P}$ and $\mathbb{P}'$ with $\hat{\mathbb{P}}(R) = 1$ if and only if*

$$\mathbb{P}(U) \leq \mathbb{P}'(N_R(U)) \quad \text{for all } U \subseteq A,$$

where $N_R(U) := \{y \in B\colon\ \exists x \in U \text{ such that } (x, y) \in R\}$.

Strassen's theorem is closely related to a celebrated result in combinatorics on perfect matchings in bipartite graphs, known as the *marriage theorem* due to Hall [33]. A *perfect matching* of a graph is a set of edges such that each vertex is incident to exactly one edge. In the literature, the perfect matchings of a graph are studied using the dimer model [43]. In our work, however, matchings will be approached from a combinatorial perspective.

The problem addressed by the marriage theorem asks whether a given bipartite graph has a perfect matching.

Theorem 1.2 (Hall's marriage theorem). *Let G be a bipartite graph with bipartition $\{A, B\}$ such that $|A| = |B|$. Then G contains a perfect matching if and only if*

$$|U| \leq |N_G(U)| \quad \text{for all } U \subseteq A,$$

where $N_G(U)$ denotes the set of vertices that are neighbors of vertices in U.

It is well-known that the marriage theorem belongs to a larger class of combinatorial theorems that are all *equivalent* to each other, in the sense that each of them can be easily derived from any of the others. See [73] for an overview of these equivalences.

Strassen's theorem belongs to this class of equivalent theorems as well, as shown by Dudley, who derives Strassen's theorem on stochastic domination from the marriage theorem [23]. Another elegant proof of Strassen's theorem for finite sets is mentioned in [58, pp. 46]. This proof, which is elaborated in e.g. [36], derives Strassen's theorm from one of the equivalent theorems, namely, Ford and Fulkerson's *max-flow min-cut theorem* [28], and shows that any method for finding maximal network flows can be used to construct the sought coupling.

1.3 Outline

Chapter 2 In chapter 2 we study the connectivity properties of the Kirchhoff forest measure. Each rooted forest partitions the vertices of the underlying graph into its constituent connected components. By considering the vertex partition resulting from a Kirchhoff forest, we obtain a distribution on the set of partitions of the vertices, the so-called *loop-erased partitioning*. The observable of interest in this chapter is the *two-point correlation function*, defined as the probability that two given vertices belong to the same block of a loop-erased partition, or equivalently, as the probability that these vertices belong to the same tree of a Kirchhoff forest.

First, the monotonicity of the two-point correlations is established, when considered as a function of the intensity parameter q, for the special case in which the underlying graph is undirected. We conjecture that this monotonicity extends to the general setting.

Next, we continue an investigation initiated in [7], asking whether the two-point correlations can detect various clustering structures in the underlying graph. Several simplistic sparse graphs with and without built-in structures are considered as examples. By studying the asymptotics of the two-point correlations as the number of vertices increases, we investigate which of the built-in structures can be detected by correctly tuning the intensity parameter q. Special emphasis is given to path graphs, for which we show that detailed control on the asymptotics of the two-point correlations can be obtained.

Chapter 3 In chapter 3 we define and study a stochastic process that is a dynamic extension of the occupation field of Wilson's algorithm. The construction of this process utilizes an extension of the Kirchhoff forest measure that was introduced by Avena and Gaudillière in [6], which couples Kirchhoff forest measures with different intensities. A complete characterization is given for the distribution of the process. In particular, it is shown that the constructed occupation field process is a piece-wise constant and monotone increasing Markov process, and that at each jump time the increment of the process is distributed as the occupation field of a single loop of a random walk loop-soup.

Chapter 4 The investigation of chapter 3 is continued in chapter 4, where the focus shifts from the occupation field of Wilson's algorithm to the configurations of removed loops from which the occupation field can be obtained. As in chapter 3,

the coupling of Kirchhoff forests from [6] is used to extend the static random loop configuration, obtained from a single application of Wilson's procedure, to a dynamic process of increasing amounts of loops.

Utilizing this dynamic loop configuration, three results are achieved in chapter 4. Firstly, a dynamic extension of the random walk loop-soup is constructed. Incidentally, this construction provides an alternative proof of the result in chapter 3 which is more insightful than the proof provided there. Secondly, a spectral decomposition of the dynamic random walk loop-soup is provided, which could aid the refinement of the spectrum estimation scheme proposed in [9]. Thirdly, by relying on Lupu's coupling of the random walk loop-soup and the discrete Gaussian free field with mass [59], a coupling is constructed that couples Gaussian free fields of all possible masses.

Chapter 5 The main contribution of chapter 5 is a novel and elementary proof of Strassen's theorem for the special case in which sets A and B are finite, that utilizes a forest reduction scheme for the coupling problem. By interpreting R as the edges of a bipartite graph with vertices $A \cup B$, the problem can be translated into a graph-theoretic setting. The main lemma of chapter 5, the so-called *subforest lemma*, shows that the relevant structure of this bipartite graph can be captured by a single forest. This lemma is used to derive Strassen's theorem for finite sets. Contrary to the results in part I, the forest reduction scheme in chapter 5 does not aim to be of practical use for efficiently constructing the sought coupling.

Further, a derivation of Strassen's theorem for finite sets from the marriage theorem is given, which establishes the equivalence of the two theorems. This derivation is an adaptation of Dudley's proof in [23].

1.4 Open problems and further research

One conjecture has already been mentioned above, and concerns the two-point correlation function, i.e. the probability that two vertices belong to the same tree of the Kirchhoff forest, which is studied in chapter 2.

- Is the function $q \mapsto \mathbb{P}(x \leftrightarrow_{\Phi_q} y)$ monotone non-decreasing on any weighted directed graph $\mathcal{G}$ for all vertices x and y?

It is shown in chapter 4 that several observables of Wilson's occupation field can be expressed in terms of the Laplacian spectrum. These observables have an advantage over the spectral observables that are employed in the Kirchhoff forests based spectrum estimation method that is proposed in [9]. Namely, the occupation field observables are expressed as a *mixture* of random variables each of which depends only on a single eigenvalue, rather than as a sum of such variables. Moreover, some of the occupation field observables depend not only on the Laplacian spectrum, but also on the Laplacian eigenvectors.

- How can the spectral observables of Wilson's occupation field be employed for a spectral estimation method, and can they be used to devise an eigenvector estimation method?

Two further open problems are related to the *coupled forest process*, which is a coupling of Kirchhoff forests of different intensities, and is described in detail in chapter 3. This interesting process is ill understood, and warrants further research. While a complete description of this process might be out of reach, the following two problems provide interesting starting points.

One starting point would be to study the loop-erased partitioning process, obtained by considering only the dynamic vertex partitioning resulting from the coupled forests.

- What is the distribution of the loop-erased partitioning process?

Even on simple geometries such as the path graphs this question would be of interest.

While the intensity parameter q can be tuned to provide a given expected number of roots, sampling a Kirchhoff forest conditioned to have exactly k roots is not as simple. Denote by τ_k the first hitting time by the coupled forest process $(\Phi_{1/t})_{t\geq 0}$ of the set of forests with k roots.

- What is the distribution of the coupled forest Φ_{1/τ_k} at this hitting time, and what is the distance[2] between its law $\mathbb{P}(\Phi_{1/\tau_k} \in \cdot)$ and the conditioned law $\mathbb{P}(\Phi_q \in \cdot \mid r(\Phi_q) = k)$?

An interesting direction in which future research on Kirchhoff forests could be taken is the study of related thermodynamic and scaling limits, which might exhibit interesting phase transitions.

[2] Any relevant notion of distance between distributions can be used, e.g. the total variation distance.

PART I

KIRCHHOFF FORESTS

CHAPTER 2

Loop-erased partitioning via parametric spanning trees

This chapter is based on the following paper: L. Avena, J. E. P. Driessen, and V. T. Koperberg. "Loop-erased partitioning via parametric spanning trees: Monotonicities & 1D-scaling". In: *Stochastic processes and their applications* 176 (2024), p. 104436.

Abstract

We consider a parametric version of the UST (Uniform Spanning Tree) measure on arbitrary directed weighted finite graphs with tuning (killing) parameter $q > 0$. This is obtained by considering the standard random weighted spanning tree on the extended graph built by adding a ghost state † and directed edges to it, of constant weights q, from any vertex of the original graph. The resulting measure corresponds to a random spanning rooted forest of the graph where the parameter q tunes the intensity of the number of trees as follows: partitions with many trees are favoured for $q > 1$, while as $q \to 0$, the standard UST of the graph is recovered. We are interested in the behaviour of the induced random partition, referred to as loop-erased partitioning, which gives a correlated cluster model, as the multiscale parameter $q \in [0, \infty)$ varies.

Emergence of giant clusters in this correlated percolation model as a function of q has been recently explored on certain dense growing graphs [7]. Herein we derive two types of results. First, we explore monotonicity properties in q of this forest measure on general graphs showing in particular some counter-intuitive subtleties in non-reversible settings where the electrical-network interpretation of the UST observables gets partially lost. Second, by analyzing two-point correlations on trees and various very sparse growing graph models, we characterize emerging macroscopic clusters, as q scales with the graph size, and derive related phase diagrams.

2.1 Rooted spanning forests, loop-erased partitioning and weighted spanning tree measures

Consider an arbitrary directed weighted finite graph $G = (V, E, w)$ on $n = |V|$ vertices where $E \subseteq \{e = (x, y) : x, y \in V\}$ stands for the edge set and $w : E \to [0, \infty)$ is a given edge-weight function. We call Random Walk (RW) associated to G the continuous-time Markov chain $X = (X_t)_{t \geq 0}$ with state space V and infinitesimal generator L given by the negative of the graph Laplacian, i.e. L is the $n \times n$ matrix with off-diagonal entries $L_{xy} = w(x, y)$ and diagonal entries $L_{xx} = -\sum_{z \in V \setminus \{x\}} w(x, z)$ guaranteeing that the entries of each row in L sum up to 0.

A *spanning rooted forest* of a graph is a union of vertex disjoint rooted trees spanning its vertex set, where we consider a rooted tree to be a collection of directed edges pointing towards the root. That is, a rooted forest F is a subset of E such that:

(i) each vertex has at most one outgoing edge in F;

(ii) if there exists a directed path in F from vertex x to vertex y, then no such path exists from y to x.

The *roots* of F are those vertices without an outgoing edge. Let $\mathcal{F}$ denote the collection of all spanning rooted forests of G.

Definition 2.1.1 (Rooted Spanning Forest of intensity q). Fix a positive parameter $q > 0$ and let Φ_q be the random variable with values in $\mathcal{F}$ with law:

$$\mathbb{P}(\Phi_q = F) = \frac{q^{r(F)} w(F)}{Z(q)}, \qquad F \in \mathcal{F}, \tag{2.1}$$

where $w(F) := \prod_{e \in F} w(e)$ stands for the forest weight, $r(F)$ denotes the number of trees (or equivalently the number of roots) in $F \in \mathcal{F}$, and $Z(q)$ is a normalizing constant referred to as the partition function. We will refer to this measure as *random spanning rooted forest* of intensity q. ■

In the unitary weight case $w \equiv 1$, when $q = 1$, this measure becomes uniform over the set of spanning rooted forests $\mathcal{F}$ and its structure has been partially analyzed in several geometrical setups in relation to random combinatorial models in statistical physics and coalescence theory, see [18, 40, 41, 44, 45, 56, 71, 72]. For any $q > 0$, Φ_q induces a randomized decomposition of a given network into blocks (corresponding to its trees) and for each block it identifies a representative node (the root of a tree). The presence of the tuning parameter q makes this object natural for exploring a network architecture in a multiscale fashion. The goal of this paper is to understand the structure of the resulting unrooted random blocks on the set of partitions $\mathcal{P}(V)$ of the vertex set V as the scaling parameter q varies. We refer to this object, defined below, as the Loop-Erased Partitioning (LEP). Its analysis has been initiated on dense graphs in the recent [7]. In this work we derive general results on the monotonicity properties of this measure (see Section 2.2.1) and then, by means of these and other properties, we perform a systematic analysis of the emergent partition on various very sparse simple growing topologies (see Section 2.2.4).

Definition 2.1.2 (Loop-Erased Partitioning (LEP) of intensity q). Given $G = (V, E, w)$, fix a positive parameter $q > 0$. We call *loop-erased partitioning of intensity* q, the random unrooted partition, denoted by Π_q, of V, with law:

$$\mathbb{P}(\Pi_q = \pi_m) = \frac{q^m \sum_{F\in\mathcal{F}:\pi(F)=\pi_m} w(F)}{Z(q)}, \qquad \pi_m \in \mathcal{P}(V),\ m \le |V|, \tag{2.2}$$

where the sum runs over the space of spanning rooted forests $\mathcal{F}$ of G and $\pi(F)$ stands for the partition of V induced by a given spanning rooted forest F where each block is determined by vertices belonging to the same tree, and m counts the number of blocks in the partition π_m. Equivalently,

$$\Pi_q := \pi(\Phi_q). \tag{2.3}$$

■

Rooted forests and spanning tree measure with uniform killing.

The rooted forest Φ_q is a natural extension of the classical UST (Uniform Spanning Tree) measure which on strongly connected graphs is readily recovered in the constant weight case $w \equiv 1$ by taking the limit of q going to zero in Eq. (2.1). Alternatively, this rooted forest Φ_q can also be seen as a measure on weighted spanning trees on the extended weighted graph obtained by adding an extra cemetery state accessible from any vertex via an edge with weight q. Under this perspective, it is clear that most results known for the UST do have a generalized analogue in the context of this rooted forest measure. For example, edges in Φ_q form a determinantal process [6] due to a version of the so-called transfer-current theorem [17], clarifying its status within negatively associated systems, see [31, 42, 68]. Due to the Kirchhoff's matrix tree theorem, the normalizing constant in Eq. (2.2) can be expressed as the characteristic polynomial of the matrix L evaluated at q, i.e.

$$Z(q) := \sum_{F\in\mathcal{F}} q^{r(F)} w(F) = \det[qI - L], \tag{2.4}$$

see e.g. [6, 19]. As far as sampling is concerned, for fixed $q > 0$, one can use the celebrated algorithm due to Wilson [77] based on loop-erased random walks. The latter is in fact a classical efficient procedure allowing to sample a rooted tree of a graph with probability proportional to its weight. Further, it is well known that the UST can be obtained from the unifying Fortuin-Kasteleyn-percolation 'super-model' by properly taking the related interaction parameter to zero, see e.g. [30]. Not surprisingly, as expressed in Lemma 2.1 below, which for simplicity we state in the unitary weight case $w \equiv 1$, the rooted forest in Eq. (2.1) can also be obtained via a similar zero-limit but by considering a proper FK-percolation with an additional cemetery state. The proof of this proposition is as in [30], see Thm. 1.23 in Sect 1.5 therein, with the parameters of the FK as specified in the statement below.

Lemma 2.1 (Rooted forest as zero-limit of extended FK-percolation). *Given an undirected simple graph $G = (V, E)$, let $G_\dagger := (V_\dagger, E_\dagger)$ be the extended graph with*

$V_\dagger := V \cup \{\dagger\}$ where $\dagger$ denotes an extra state, $E_\dagger := E \cup \bar{E}$ with $\bar{E} := \{(x,\dagger) : x \in V\}$. Consider the generalized FK-percolation on $G_\dagger$ with parameter $\lambda > 0$ and vector of weights $\vec{p} = (p_e)_{e\in E_\dagger}$ such that $p_e = p \in (0,1)$ if $e \in E$ and $p_e = \gamma > 0$ for $e \in \bar{E}$, that is, the following measure on spanning subgraphs of $G_\dagger$ seen as collection of edges in $\mathcal{G} := 2^{E_\dagger}$:

$$\mathbb{P}(FK = H) = \frac{\lambda^{k(H)} \prod_{e\in E} p^{\mathbf{1}\{e\in H\}}(1-p)^{1-\mathbf{1}\{e\in H\}} \prod_{e\in \bar{E}} \gamma^{\mathbf{1}\{e\in H\}}(1-\gamma)^{1-\mathbf{1}\{e\in H\}}}{Z(\lambda,\vec{p})}, \quad (2.5)$$

for $H \in \mathcal{G}$, with $k(H)$ counting the number of connected components of the graph $(V_\dagger, H)$ and $Z(\lambda,\vec{p})$ being a normalizing constant. Assume that $\vec{p}$ is a function of λ such that, as $\lambda \to 0$, $\gamma = \gamma(\lambda) \to 0$, $p = p(\lambda) \to 0$ and $\gamma(\lambda)/p(\lambda) \to q \in (0,\infty)$. Then as λ goes to zero, the law in Eq. (2.5) (projected onto subgraphs of G) degenerates into the law of the random rooted forest Φ_q in Eq. (2.1) with unitary weights.

If the UST represents the "static global random backbone" of a given network, the forest process $(\Phi_q)_{q>0}$ can be seen as its "mesoscopic and dynamic" analogue where the notion of locality is captured parametrically by what the RW sees on time-scale $1/q$. As such, it naturally leads to dynamic multiscale approaches (see [6, Thm.2] for an associated coalescence-fragmentation process), and new structures and questions which do not make sense within the more restrictive global and static UST context.

Applications of rooted forest measure and LEP:

In a series of recent works [1, 5, 6] some general properties of the rooted forest measure have been explored. For example, the roots [6, Prop.2.2] in Φ_q form a determinantal point process with kernel related to the RW Green's function, that is: for any $A \subset V$

$$\mathbb{P}(A \text{ is in the set of roots}) = \det[K_q]_A, \quad (2.6)$$

with $[K_q]_A$ being the restriction of the matrix $K_q := q(qI - L)^{-1}$ to the set of indices in A. The number of roots (or trees, or blocks in Π_q) is distributed as the sum of n Bernoulli random variables with success probabilities $\frac{q}{q+\lambda_i}$, for $i \le n$, with the λ_i's being the eigenvalues of $-L$, or their real parts, see e.g. [6, Prop. 2.1]. Its mean number is monotonically increasing in q. Further, these roots turn out to be well-distributed in the given network [6, Thm.1] and, conditional on the induced partition, their joint law is determined by the stationary measures of the random walk X restricted to each block of the underlying partition [6, Prop.2.3]. These and other features of the LEP have been recently exploited to build novel algorithms for the following different applications in data science: multiresolution scheme, wavelets basis and filters for signal processing on graphs [2, 69, 70], estimate traces of discrete Laplacians and other diagonally dominant matrices [10], network renormalization [1, 3], centrality measures [19] and statistical learning [8]. These applications give further motivations to explore this LEP in more detail. Let us also stress that on general graphs or certain specific settings such as the integers, it would be of interest to study the LEP in connection to other random partitions and a natural line of investigation would be to study its intriguing dynamical structure [6, see Thm.2 and Sect 2.2] within the theory of coalescence-fragmentation processes.

Other forest measures:

To conclude this introduction let us clarify that this *rooted* forest Φ_q should not be confused with other forest measures that have been receiving a large amount of attention in the literature in relation to universality classes in statistical physics and to negatively correlated systems. In particular, when taking the (weak) infinite volume limit of the UST on d-dimensional lattices for $d > 4$ (and other transitive settings), depending on the boundary condition procedure when approaching the limit, the resulting measure concentrates on *unrooted* forests referred to as wired or free spanning forests, see e.g. [13, 14, 37, 38, 67] and references therein. On finite graphs another natural extension of the UST is obtained when considering the uniform measure on unrooted spanning forests. Properties of this other fascinating forest measure have been investigated in [12] and recently in [11].

Results overview and paper structure:

The statements of our main results are organized in Section 2.2 divided into the following four subsections.

Monotonicities on arbitrary graphs: We start in Subsection 2.2.1 by presenting results regarding the monotonicity properties of the rooted forest measure in the intensity parameter q. In particular: in Lemma 2.2 we state a general formula to characterize q-monotone events; in Theorem 2.3 we establish monotonicities on *undirected graphs* for both the edge process and the 2-point correlation function, which will later be analyzed in different graph settings to study the emergent partition; and in Remark 2.4 we then discuss subtle issues and counterexamples when trying to extend the latter results to a general non-reversible setting.

Connectivity function on trees: In Section 2.2.2 we look at general weighted directed trees. We state that the connectivity function is still monotone on this directed sub-class of graphs, Theorem 2.5, and we present a related inclusion-exclusion reduction formula, see Theorem 2.6.

$q-$scaling on growing segments: Section 2.2.3 is devoted to a detailed analysis of the LEP on the first n integers where equipartitions are favored. Formulas for the partition function are first derived in Theorem 2.7, and extended to a ring, Corollary 2.7.1. Theorem 2.8 gives a recursive representation of the pairwise correlation in terms of reduced partition functions and offers bounds in terms of the corresponding RW on the infinite line. The subsequent Corollary 2.8.1 shows explicit bulk and boundary asymptotics.

Emergence of giants on modular growing tree-like structures: Section 2.2.4 is then devoted to the exploration of the emergent blocks and detection of simple modular structures in certain insightful models for the proper $q-$scales. In particular, Proposition 2.10 and Theorem 2.11 look at a star graph without and with a community structure, respectively. Proposition 2.12 and Theorem 2.13 show similar analysis on finite trees with different weighted structures in which for different magnitudes of q different layers are detected. Finally, in Theorem 2.9

we consider asymptotic detection in a bottleneck graph with two variable-size connected complete subgraphs by combining the results on the segment, after suitable contraction, and those for the mean-field case obtained in [7].

All proofs are given in Section 2.3. Technical short lemmas are provided in the appendices.

2.2 Results: monotonicities & emergent partition

2.2.1 Monotonicity & connectivity function

A notoriously difficult issue for most of the measures that can be obtained from FK-percolation, is to establish monotonicity properties as a function of the involved parameters. The following lemma, which is reminiscent of Russo's pivotality formula in percolation models [15, 29], offers a general characterization of monotone events w.r.t. Φ_q as a function of q.

Lemma 2.2 (Monotone events for the rooted forest on arbitrary networks). *Let $G = (V, E, w)$ be a weighted directed graph, and let $r_q := r(\Phi_q)$ be the number of roots of the random rooted forest Φ_q. Then, for any set of rooted forests $\mathcal{H} \subseteq \mathcal{F}$, it holds that the derivative w.r.t q of the probability of the event $\Phi_q \in \mathcal{H}$ is given by*

$$\frac{d}{dq}\mathbb{P}(\Phi_q \in \mathcal{H}) = \tfrac{1}{q}\mathbb{P}(\Phi_q \in \mathcal{H})\big(\mathbb{E}[r_q \mid \Phi_q \in \mathcal{H}] - \mathbb{E}[r_q]\big). \tag{2.7}$$

This statement, which is proven in Section 2.3.1, is a special case of the more general fact that for Gibbsian measures μ_β with Hamiltonian H and inverse-temperature β it holds that $\frac{\mathrm{d}}{\mathrm{d}\beta}\int X \,\mathrm{d}\mu_\beta = -\int HX \,\mathrm{d}\mu_\beta$. The main advantage of Lemma 2.2 compared to this more general statement is the geometrical interpretation of the right hand side in terms of root numbers. It shows that monotone events in q are those for which the difference $\mathbb{E}[r_q \mid \Phi_q \in \mathcal{H}] - \mathbb{E}[r_q]$ has a constant sign as q varies. In practice it might be not straightforward to check the sign of this difference, since it requires control on the conditional distribution of r_q. Still, for specific events we believe this statement can be of great help, of which we give an example in the proof of Theorem 2.3. We also mention that in [6] a coupled version of the forest[1] is constructed by means of an algorithm allowing to sample an entire forest trajectory $(\Phi_q)_{q\in[0,\infty)}$. Yet, this coupling is monotone only in mean, but not trajectory-wise, hence this coupling is not useful to characterize monotone events.

As anticipated, our main interest within this work is to explore monotonicity properties of this loop-erased partitioning and its detailed structure on trees and nearly-one-dimensional geometries. To do so, we will mainly analyze 2-point correlations associated to Π_q, which we introduce next. For a pair of distinct vertices $x, y \in V$, consider the event that these vertices belong to different blocks in Π_q. That is, the

[1] This coupling corresponds to an explicit Markovian coalescence-fragmentation process with values in $\mathcal{F}$ in which coalescence of trees is dominant but whenever the underlying building RW produces a loop, a tree gets fragmented into subtrees, see [6, see Thm.2 and Sect2.2].

event

$$\{B_q(x) \neq B_q(y)\} := \{x \text{ and } y \text{ are in different blocks of } \Pi_q\},$$

where $B_q(z)$ stands for the block in Π_q containing $z \in V$.

Definition 2.2.1 (2-point correlation or connectivity function). For given $q > 0$ and G, and any pair $x, y \in V$, we call *connectivity function* the following probability:

$$\begin{aligned} U_q(x,y) :=& \mathbb{P}(B_q(x) \neq B_q(y)) \\ &= \sum_{\gamma} \mathbb{P}_x^{LE_q}(\gamma)\mathbb{P}_y(\tau_\gamma > \tau_q) \end{aligned} \tag{2.8}$$

where τ_q denotes an independent exponential random variable of rate q, $\mathbb{P}_z$ and $\mathbb{P}_z^{LE_q}$ stand for the laws of the RW X and the corresponding loop-erased RW killed at rate q, respectively, starting from $z \in V$. Further, the above sum runs over all possible self-avoiding paths γ starting at x and $\tau_\gamma := \inf\{t \geq 0 : X_t \cap \gamma \neq \varnothing\}$ is the random walk hitting time of the set of vertices in γ. ■

The representation in Eq. (2.8) is a consequence of Wilson's sampling procedure and it holds true since, remarkably, this algorithm is exchangeable with respect to the starting point of each loop-erased random walk launched along the algorithm steps [77]. Furthermore, we notice that, as for any generic random partition of V, such a connectivity function defines a distance on the vertex set. This specific metric $U_q(x,y)$ can be interpreted as an affinity measure capturing how densely connected vertices x and y are in the graph G.

Our second general result, Theorem 2.3, further explores monotonicities in q when considering undirected networks. Since spanning rooted forests impose a directionality on its edges, it is convenient to interpret an undirected graph as a symmetric directed graph with a symmetric weight function, $w(x,y) = w(y,x)$ for $(x,y) \in E$. For these symmetric graphs Theorem 2.3 states that the "unoriented" edge process, see (2.9), as well as the 2-point correlations, see (2.10), are both monontone in q. To state the result about the edge process, we will use the following notation. For a directed edge $e = (x,y)$ write $e^- = (y,x)$ to denote its reversed edge, and let $\{\pm A \subseteq \Phi_q\} = \bigcap_{e \in A}(\{e \in \Phi_q\} \cup \{e^- \in \Phi_q\})$ denote the event that for each edge $e \in A$ either e or e^- is present in the random rooted forest Φ_q.

Theorem 2.3 (Monotonicity of edges and 2-point correlations on undirected networks). *Consider a symmetric weighted directed graph $G = (V, E, w)$ and the rooted forest Φ_q on G for $q \in [0, \infty)$. Let $A \subseteq E$ be a set of directed edges, then the function*

$$q \mapsto \mathbb{P}(\pm A \subseteq \Phi_q) \tag{2.9}$$

is monotone non-increasing. Furthermore, for any distinct $x, y \in V$, the function

$$q \mapsto U_q(x,y) \tag{2.10}$$

is continuous and non-decreasing with $U_0(x,y) = 0$ and $\lim_{q\to\infty} U_q(x,y) = 1$.

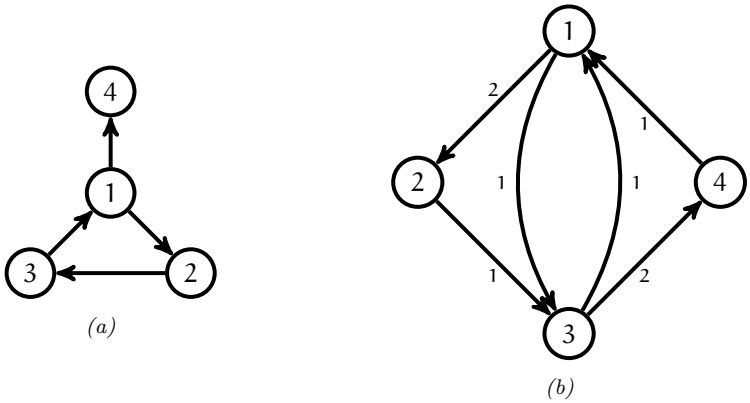


Figure 2.1: Two directed graphs for which the edge monotonicity (2.9) *does* not *hold.*

The monotonicities presented in Theorem 2.3 on undirected graphs are not too surprising, as this result could be derived from the well-known negative association result of Feder and Mihail for balanced matroids, see [60, Theorem 4.6 and Lemma 10.3]. This becomes apparent when we interpret the killing rate as edge weights on an extended graph (as in Lemma 2.1), and in turn interpret all edge weights as multi-edges, which is straightforward for integer edge weights, but can also be achieved for real valued edge weights via rational approximation and up-scaling of rational edge weights into integers.

Remark 2.4 (Beyond reversibility & main open problem). The proof of Theorem 2.3 presented in Section 2.3.1 will not use the result of Feder and Mihail, and will instead showcase Lemma 2.2, which could in principle be applied in a directed setting. The Feder and Mihail result, in contrast, fails in directed setting, as rooted trees do not exhibit the same matroid basis structure of undirected trees. Our proof will exploit the undirectedness assumption, but we believe such monotonicity to be valid in greater generality, though this remains a delicate open problem.

The edge monotonicity of (2.9) should be extendable to reversible graphs, i.e. strongly connected graphs on which the RW is reversible, since reversible graphs can be identified with symmetric graphs with inhomogeneous killing, to which the proofs presented in Section 2.3.1 are still applicable. Moreover, (2.10) might very well hold for all weighted directed graphs. As will become clear in the proof of Theorem 2.3, the monotonicity of the (unoriented) edge process in (2.9) implies in particular the monotonicity of the connectivity function in (2.10). However, these two monotonicities are not equivalent. As the counterexamples depicted in Figure 2.1 in fact show, it is not difficult to find non-reversible graphs for which the monotonicity of the edges in (2.9) fails whereas the one in (2.10) does still hold true. Indeed, for the unweighted graph in Figure 2.1a it holds that

$$\mathbb{P}(\pm\{(1,2)\} \subseteq \Phi_q) = \mathbb{P}((1,2) \in \Phi_q) = \frac{2q^2 + q^3}{q + 5q^2 + 4q^3 + q^4},$$

which is increasing for $q < \sqrt{3} - 1$. This failure of monotonicity can be explained

heuristically with the help of Lemma 2.2. For q very small the expected number of roots is close to one. Conditioning on edge $(1,2)$ being present forces the random forest to have at least two roots, so then the conditioned expectation of the number of roots will be larger then the unconditional expectation. The above heuristic uses that the graph is not strongly connected. In strongly connected graphs, conditioning on unoriented edges being present cannot increase the minimal number of roots in a random forest. Still, Figure 2.1b shows a strongly connected weighted graph with

$$\mathbb{P}(\pm\{(1,3)\} \subseteq \Phi_q) = \frac{6q + 8q^2 + 2q^3}{18q + 21q^2 + 8q^3 + q^4},$$

which is increasing for $q < \sqrt{2} - 1$.

Unlike the subtle edge monotonicity in (2.9), we could not find examples where the one for the 2-point correlation (2.10) fails and in fact, we conjecture the latter to hold true in general but proving it remains an open problem. A first careful attempt to settle this conjecture in a non-reversible setting is offered in Theorem 2.5, where it is shown that (2.10) also holds at least on arbitrary weighted directed trees. In particular, on trees that are not strongly connected and hence non-reversible. It should be noted that even trees that are not strongly connected do satisfy the cycle condition for reversibility [62, p. 307], by virtue of having no non-trivial cycles. So, in some sense such trees are still close to be reversible.

2.2.2 Two-point correlation on trees

We start here to discuss results specific to trees, which are weighted directed graphs in which for any pair of vertices there exists a unique undirected path between them. Let us notice that in this setup, the analysis is facilitated by the absence of cycles. In general, the mapping from $\mathcal{F}$ to rooted partitions is not injective, while on trees this is the case.

In the constant weight case $w \equiv 1$, for a partition into $m \leq |V|$ blocks $\pi_m = \{B_1, B_2, \ldots, B_m\} \in \mathcal{P}(V)$, the measure in Equation (2.2) reads as

$$\mathbb{P}(\Pi_q = \pi_m) = \frac{q^m \prod_{i=1}^m |B_i|}{Z(q)},$$

from which we see that, for a given q, it concentrates on partitions where the block sizes tend to be of the same order. In this sense equipartitions are favored.

The first result in this tree specific setting extends the monotonicity of the 2-point correlation, as expressed in Theorem 2.3, to a specific weighted directed setting.

Theorem 2.5 (Monotonicity of 2-point correlations on trees). *If $G = (V, E, w)$ is a weighted directed tree, then for all $x, y \in V$ the function*

$$q \mapsto U_q(x, y)$$

is monotone non-decreasing.

Next we derive a representation of the 2-point correlation on arbitrary trees, in terms of reduced partition functions over subtrees.[2] To avoid confusion, in each statement in the sequel we will add proper indices to the partition and connectivity functions specifying the considered graph. The distance $d(x, y)$ between two vertices x and y will refer to the unweighted shortest path distance, i.e. the minimum number of edges on an undirected path between the two vertices.

Theorem 2.6 (Inclusion-exclusion for 2-point-correlation on trees). *Let $G = (V, E, w)$ be a weighted directed tree. Fix $x, y \in V$ with $d(x, y) = d$ and let $(z_i)_{i=0}^{d}$ be the unique undirected path with $z_0 = x$ and $z_d = y$. For a subset $I \subseteq [d]$ let G_I denote the graph obtained by removing all edges between z_{i-1} and z_i from G for all $i \in I$. Denote the $|I| + 1$ connected components of G_I by $G_I^1, \dots, G_I^{|I|+1}$. Then, for every $q > 0$, the following representation is valid*

$$U_q^{(G)}(x, y) = \frac{1}{Z_G(q)} \left(\sum_{k=1}^{d} (-1)^{k+1} \sum_{I \in \binom{[d]}{k}} \prod_{i=1}^{k+1} Z_{G_I^i}(q) \right). \tag{2.11}$$

Here $\binom{[d]}{k}$ denotes the collection of k-element subsets of $[d]$.

In particular for x, y such that $d(x, y) = 1$:

$$U_q(x, y) = \frac{Z_x(q) Z_y(q)}{Z_G(q)}, \tag{2.12}$$

where $Z_x(q)$ and $Z_y(q)$ denote the partition functions of the two connected components of the graph obtained by removing the edges between x and y.

2.2.3 Integer partitioning: analysis on lines and rings

In what follows we denote by $PG_n := \mathbb{Z} \cap [1, n]$ the (undirected and unweighted) path-graph constituted by the first n integers and by CG_n the cycle-graph on n vertices (i.e. the one dimensional discrete torus).

Theorem 2.7 (Partition function of path-graphs). *The partition function in* (2.4) *of PG_n can be expressed in the following ways:*

$$\begin{aligned} Z_{PG_n}(q) &= \sum_{k=1}^{n} \binom{n+k-1}{2k-1} q^k && (2.13) \\ &= \prod_{k=1}^{n} \left(q + 2 - 2\cos\left(\tfrac{\pi(n-k)}{n}\right)\right) && (2.14) \\ &= \frac{q\left(q+2+\sqrt{q^2+4q}\right)^n - q\left(q+2-\sqrt{q^2+4q}\right)^n}{2^n\sqrt{q^2+4q}} && (2.15) \\ &= qT_{n-1}(\tfrac{q}{2}+1). && (2.16) \end{aligned}$$

Here T_{n-1} denotes the $n-1$-th degree Chebyshev polynomial of the second kind.

[2] This type of reduction is due to the well-known spatial Markov property for UST-like measures (see Proposition A.2).

As can be appreciated in the proof, the above different representations reflect different computational methods suited for the random forest. We notice that for $q = 1$ evaluating this partition function corresponds to counting the number of rooted forests of the path-graph, as previously derived in [18].

One of the messages of this paper is that having an explicit characterization of a simple given geometry can be useful to derive information on some more involved geometry. The next corollary shows one such very simple instance by expressing the partition function on the torus in terms of partition functions of the simpler path-graph.

Corollary 2.7.1 (Partition function of cycle-graphs). *The partition function of CG_n is given by*

$$Z_{CG_n}(q) = Z_{PG_n}(q) + \tfrac{2}{q}\left[Z_{PG_n}(q) - Z_{PG_{n-1}}(q)\right] - 2 \tag{2.17}$$

$$= \sum_{k=1}^{n} \left(\binom{n+k}{2k} + \binom{n+k-1}{2k} \right) q^k. \tag{2.18}$$

The following result uses the notations $[n] = \{1, 2, \dots, n\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

Theorem 2.8 (Bounds for correlations on path-graph via random walk on $\mathbb{Z}$). *Let $x, y \in [n]$ be two vertices in PG_n at distance $d := y - x > 0$. Then, for any $q > 0$, the 2-point correlation between x and y is given by*

$$U_q^{(PG_n)}(x,y) = 1 - \frac{Z_{PG_{n-d}}(q)}{Z_{PG_n}(q)} - \frac{d\left[Z_{PG_x}(q) - Z_{PG_{x-1}}(q)\right]\left[Z_{PG_{n-y+1}}(q) - Z_{PG_{n-y}}(q)\right]}{qZ_{PG_n}(q)}. \tag{2.19}$$

Moreover, by denoting with $S = (S_m)_{m\in\mathbb{N}_0}$ the discrete-time simple random walk on $\mathbb{Z}$ starting at 0, the following bounds are satisfied

$$\left(1 - \left(\tfrac{2}{2+q}\right)^m\right)^2 \left(2\mathbb{P}(|S_m| < \tfrac{d}{2}) - 1\right)^2 \leq U_q^{(PG_n)}(x,y) \leq 1 - \mathbb{P}(|S_m| > d)\left(\tfrac{2}{2+q}\right)^m, \tag{2.20}$$

where the upper bound is valid for any $m \in \mathbb{N}$, while the lower bound holds for m such that $\mathbb{P}(|S_m| < \frac{d}{2}) \geq \frac{1}{2}$.

From the above statement, due to the diffusive behavior of the simple random walk S, it is clear that the correlation function between two points in a segment is non-degenerate when q_n scales with the inverse square distance between the two points. The next corollary makes this statement precise and shows that boundary effects emerge neatly from the asymptotic analysis.

Corollary 2.8.1 (Non-degenerate scaling and asymptotic boundary effects). *For each $n \in \mathbb{N}$ let x_n and y_n be vertices in PG_n. Let d_n denote the distance between these vertices and let $(q_n)_{n\in\mathbb{N}}$ be a monotone sequence of positive parameters. Then, if the limit $\lim_{n\to\infty} U_{q_n}^{(PG_n)}(x_n, y_n)$ exists, it holds that*

$$\lim_{n\to\infty} U_{q_n}^{(PG_n)}(x_n, y_n) \in (0,1) \text{ if and only if } q_n = \tfrac{c}{d_n^2} + o(\tfrac{1}{d_n^2}) \text{ for some constant } c > 0.$$

In particular, fix $\delta > 0$ and let $(\zeta_n)_{n\in\mathbb{N}}$ be a sequence such that $\zeta_n \in [\delta\sqrt{n}, n-\delta\sqrt{n}]$ for large enough n. Set $x_n = \zeta_n - \delta\sqrt{n} + o(\sqrt{n})$, $y_n = \zeta_n + \delta\sqrt{n} + o(\sqrt{n})$ and $q_n \sim \frac{1}{d_n^2}$, then the following two limits, distinguishing between the bulk and near the boundaries, are possible:

$$\lim_{n\to\infty} U_q^{(PG_n)}(x_n, y_n) = \begin{cases} 1 - \frac{3}{2e} & \text{if } \zeta_n = \omega(\sqrt{n}) \text{ and } \zeta_n = n - \omega(\sqrt{n}) \\ 1 - \frac{3}{2e} - \frac{1}{2}e^{-\frac{\alpha}{\delta}} & \text{if } \exists \alpha \geq \delta \text{ s.t. } \begin{cases} \zeta_n = \alpha\sqrt{n} + o(\sqrt{n}) \\ \quad \text{or} \\ \zeta_n = n - \alpha\sqrt{n} + o(\sqrt{n}) \end{cases} \end{cases} \tag{2.21}$$

In the above statement we computed the exact asymptotics only when the distance of the two vertices scales as the square root of n. Similar exact computations can be derived for other choices of the magnitude of this distance. We refer the interested reader to [52] for analogous statements in the cases when d_n stays of order one or diverges linearly. In particular, we note that *giants* (i.e. blocks of order $|V|$) appear at scale $q_n \sim d_n^{-2}$ and a unique giant emerges as soon as $q_n = o(n^{-2})$.

2.2.4 Detecting modular structures

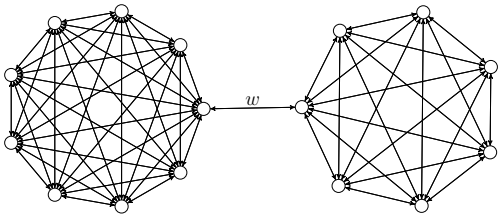
We collect here a number of simple statements of similar flavour aiming to illustrate that in tree-like graphs the emergence of giants and other modular structures can be detected with high probability by properly tuning q. Figures 2.2, 2.3 and 2.4 give a graphical overview of the main results in this section, which are given in Theorems 2.9, 2.11 and 2.13.

We start by showing with an illustrative example how the analysis on trees presented here and that on complete graphs, pursued in [7], can be combined to obtain results on mixed geometrical setups. The resulting regimes are summarized in the phase diagram in Figure 2.2b.

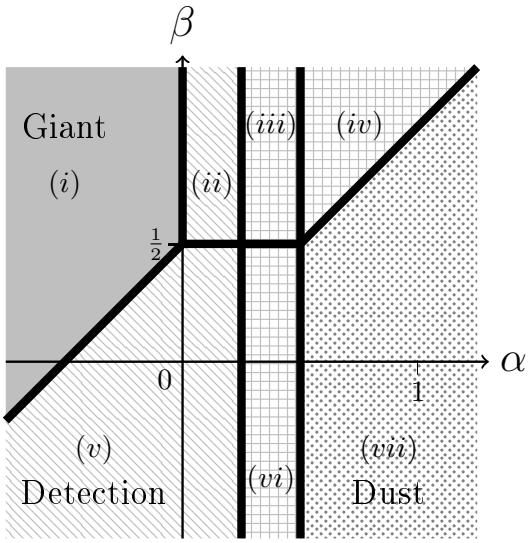
Theorem 2.9 (Detection of cliques in a bottleneck graph). *Let $BG_{n,m}$ be a bottleneck (two-cluster) graph. That is, an undirected graph consisting of two disjoint cliques C_1, C_2 on n and m vertices, respectively, that are connected via a single bridge edge, as depicted in Figure 2.2a. Equip $BG_{n,m}$ with a weight function that assigns weight w to the bridge and weight 1 to all other edges. Then its partition function is given by*

$$Z(q) = q\left(q(q+n)(q+m) + w(q+1)(2q+n+m)\right)(q+n)^{n-2}(q+m)^{m-2}. \tag{2.22}$$

Further, set $q = q_n > 0$ and let $w = w_n$ and $m = m_n$ depend on n where $n \geq m$. Denote by b, b' the two vertices incident to the bridge, by x, x' two vertices that both belong to the clique C_i containing b, and by y a vertex in the clique containing b'.



(a) A bottleneck graph with bridge weight w and two cliques of size $n = 9$ and $m = 7$.



(b) Phase diagram for the bottleneck graph, with $q = n^\alpha$, $w = n^\beta$ and $m = \sqrt{n}$. Regions (ii) and (v) are the regimes where the LEP detects the community structure. For each of the regions the following event occurs with high probability: (i) One single tree; (ii) Two trees on $n + 1$ and $\sqrt{n}$ vertices, with the large tree containing both bridge vertices; (iii) One tree consists of the n vertices in the largest clique with the bridge vertex from the small clique, while the other vertices in the small clique are isolated; (iv) Both bridge vertices are connected, and all others are isolated; (v) Two trees with n and $\sqrt{n}$ vertices, while the bridge edge is absent; (vi) One tree with all n vertices in the largest clique, while the $\sqrt{n}$ vertices in the small clique are isolated; (vii) $n + \sqrt{n}$ isolated vertices.

Figure 2.2: Summary of the results in Theorem 2.9.

Then as $n \to \infty$ it holds for the 2-point correlation between these vertices that

$$U_q(x,x') \to \begin{cases} 0 & \text{if } q = o(\sqrt{|C_i|}) \\ 1 & \text{if } q = \omega(\sqrt{|C_i|}) \end{cases} \tag{2.23}$$

$$U_q(b,b') \to \begin{cases} 0 & \text{if } q = o(\frac{w}{m}) \text{ or } (q = o(w),\ w = \omega(m)) \\ 1 & \text{if } q = \omega(w) \text{ or } (q = \omega(\frac{w}{m}),\ w = o(m)) \end{cases} \tag{2.24}$$

$$U_q(b,x) \to \begin{cases} 0 & \text{if } \begin{cases} q = o(1) \text{ or } (q = o(\sqrt{|C_i|}),\ w = o(m)) \\ \text{or } (q = o(\sqrt{|C_i|}),\ m = o(n)), \end{cases} \\ \frac{c}{1+c} & \text{if } \begin{cases} q = \omega(1),\ q = o(\sqrt{|C_i|}),\ w = \omega(m), \\ |C_i| = n,\ m \sim cn \text{ with } c \in (0,1] \end{cases} \\ \frac{1}{1+c} & \text{if } \begin{cases} q = \omega(1),\ q = o(\sqrt{|C_i|}),\ w = \omega(m), \\ |C_i| = m,\ m \sim cn \text{ with } c \in (0,1] \end{cases} \\ 1 & \text{if } q = \omega(\sqrt{|C_i|}) \end{cases} \tag{2.25}$$

$$U_q(x,y) \to \begin{cases} 0 & \text{if } q = o(1),\ q = o(\frac{w}{m}) \\ 1 & \text{if } q = \omega(1) \text{ or } (q = o(1),\ q = \omega(\frac{w}{m})). \end{cases} \tag{2.26}$$

In Proposition 2.10 we make precise how q scales on a given large star graph with homogeneous edge weights.

Proposition 2.10 (Connectivity function on a homogeneous star graph). *Let SG_n denote the star graph on n vertices, i.e. SG_n is an undirected tree consisting of a single center vertex c that is adjacent to $n-1$ leaves. Let x, y be two distinct leaves and equip SG_n with a uniform weight function that assigns weight w to all edges. Given $q > 0$,*

$$U_q(c,x) = \frac{q(q+(n-1)w)}{(q+w)(q+nw)} \tag{2.27}$$

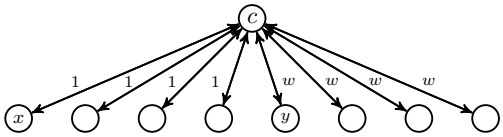
$$U_q(x,y) = \frac{q(q^2+(n+2)wq+2(n-1)w^2)}{(q+w)^2(q+nw)}, \tag{2.28}$$

which implies that $q \mapsto U_q(c,x)$ and $q \mapsto U_q(x,y)$ are strictly concave.

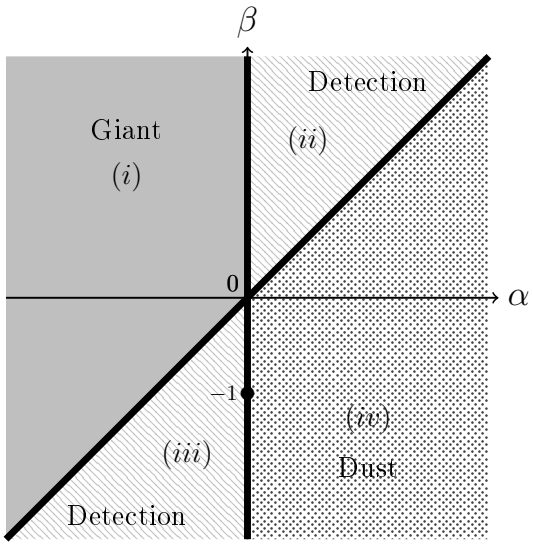
Let $q_n = \bar{q}n^\alpha$ and $w_n = \bar{w}n^\beta$ with $\alpha, \beta \in \mathbb{R}$ and $\bar{q}, \bar{w} \in (0,\infty)$. Then

$$\begin{aligned} \lim_{n\to\infty} U_{q_n}^{(SG_n)}(c,x) &= \begin{cases} 1 & \alpha > \beta \\ \frac{\bar{q}}{\bar{q}+\bar{w}} & \alpha = \beta \\ 0 & \alpha < \beta \end{cases}, \\ \lim_{n\to\infty} U_{q_n}^{(SG_n)}(x,y) &= \begin{cases} 1 & \alpha > \beta \\ \frac{\bar{q}(\bar{q}+2\bar{w}^2)}{(\bar{q}+\bar{w})^2} & \alpha = \beta \\ 0 & \alpha < \beta \end{cases} \end{aligned} \tag{2.29}$$

We see that the critical phase for the appearance of a giant is when $\alpha = \beta$. In this critical case the resulting connected subtree can be thought of as a star whose center has offspring distribution of parameter $\bar{q}/(\bar{q}+\bar{w})$.



(a) A community star graph with $n = 9$ vertices, with center vertex c and $k = 4$ vertices in the weight-1 community. The remaining four vertices belong to the weight-w community.



(b) Phase diagram for the community star graph, with $q = n^{\alpha}$ and $w = n^{\beta}$. For each of the regions the following event occurs with high probability: (i) One single tree; (ii) $k+1$ trees, all k vertices incident to a weight 1 edge are isolated, while the remaining vertices form a single tree; (iii) $n-k$ trees, all $n-k-1$ vertices incident to a weight w edge are isolated, while the remaining vertices form a single tree; (iv) n isolated vertices. The exact limit values of the correlations along the bold lines, i.e. in the non-degenerate regimes, can be found in Theorem 2.11.

Figure 2.3: Summary of the results in Theorem 2.11.

The following statement clarifies how q should be scaled in a non-homogeneous star, to detect an implanted sub-module of leaves that are more densely connected to the center. Figure 2.3b offers a graphical representation of Theorem 2.11.

Theorem 2.11 (Asymptotic detection in a star graph with two communities). *Let $CSG_{n,k}$ denote the community star graph on n vertices, which is a star graph on n vertices equipped with an inhomogeneous weight function, that assigns weight 1 to k edges and weight w to the remaining $n-k-1$ edges, as depicted in Figure 2.3a. Let c denote the center vertex, x, y vertices incident to an edge with weight 1 and w, respectively. For $\alpha, \beta \in \mathbb{R}$ take $q_n = n^\alpha$, $w_n = n^\beta$ and k constant. Then*

$$\lim_{n\to\infty} U_{q_n}^{(CSG_{n,k})}(c,y) = \begin{cases} 0 & \text{if } \alpha < \beta \\ \frac{1}{2} & \text{if } \alpha = \beta \\ 1 & \text{if } \alpha > \beta \end{cases},$$
$$\lim_{n\to\infty} U_{q_n}^{(CSG_{n,k})}(c,x) = \begin{cases} 0 & \text{if } \alpha < 0 \\ \frac{1}{2} & \text{if } \alpha = 0,\ \beta > -1 \\ \frac{k+3}{2k+8} & \text{if } \alpha = 0,\ \beta = -1 \\ \frac{k+1}{2k+4} & \text{if } \alpha = 0,\ \beta < -1 \\ 1 & \text{if } \alpha > 0 \end{cases} \tag{2.30}$$

The next two statements show similar detections on trees of different flavours.

Proposition 2.12 (Asymptotic correlation in undirected trees with bounded number of vertices). *Let $G = (V, E)$ be an undirected tree. For each $k \in \mathbb{N}$ let $w_k : E \to (0, \infty)$ be a symmetric edge weight function and $q_k > 0$ an intensity parameter. Write $G_k = (V, E, w_k)$ to denote the weighted graph obtained by equiping G with w_k. Assume that for each edge $e \in E$ the limit $\lim_{k\to\infty} \frac{w_k(e)}{q_k}$ exists in $[0, \infty]$. Fix two adjacent vertices $x, y \in V$. Then, as $k \to \infty$ it holds that*

$$U_{q_k}^{(G_k)}(x,y) \to \begin{cases} 0 & \text{if } q_k = o(w_k(x,y)) \\ 1 & \text{if } q_k = \omega(w_k(x,y)). \end{cases} \tag{2.31}$$

The following theorem holds for a specific class of undirected weighted trees that will be called 'hierarchical trees'. In these trees one vertex is specified as *ancestor* vertex. The *height* or *generation* of a vertex or edge is its distance to the ancestor. A *hierarchical tree* is a tree with edge weights $w : E \to [0, \infty)$ satisfying the following two properties:

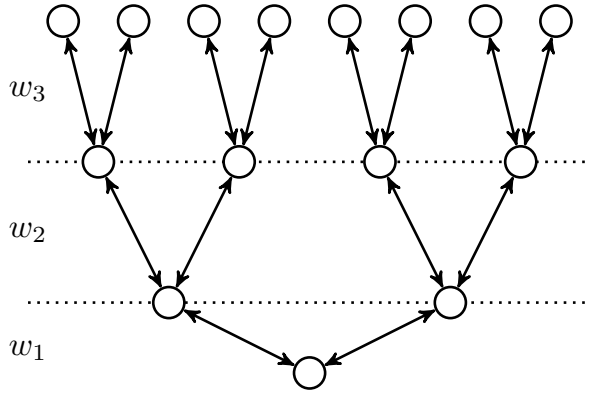
(*i*) if $e, e' \in E$ are edges in the same generation of the tree, then $w(e) = w(e')$;

(*ii*) if $e_i, e_j \in E$ are edges in generations i and j with $i < j$, respectively, then $w(e_i) \leq w(e_j)$.

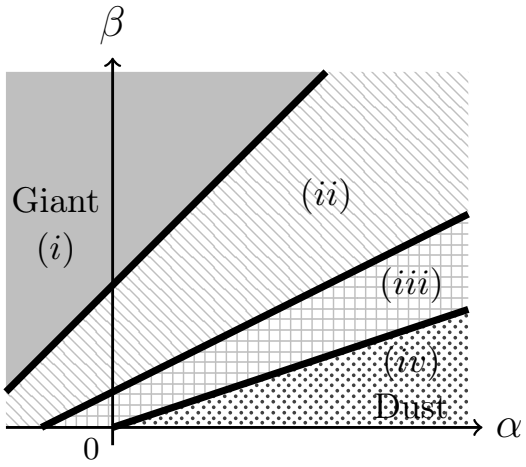
So, edges further from the ancestor of the hierarchical tree have more weight.

The *height of the tree* is the maximal height of its vertices. If x is a vertex at height h and y is a neighbor of x at height $k-1$, then we call x a *child* of y and y the *parent* of x. If each vertex with height less than the height of the tree has exactly d children, then we call the tree *complete d-ary*. The *ancestry* of a vertex is the unique

path from the vertex to the ancestor (including the vertex itself). A depiction of a complete d-ary hierarchical tree is given in Figure 2.4a.



(a) A complete binary tree of height $h = 3$ with hierarchical edge weights. Each edge in generation i has weight w_i and these weights satisfy $w_1 \leq w_2 \leq w_3$.



(b) Phase diagram for the complete d-ary hierarchical tree of height $h = 3$, with $d = n$, $q = n^\alpha$ and j-th generation edge weights $w_j = n^{j\beta}$ for $\beta \geq 0$. For each of the regions the following event occurs with high probability: (i) One single tree; (ii) All 2nd and 3rd generation edges are present, while all 1st generation edges are absent; (iii) All 3rd generation edges are present, while all 1st and 2nd generation edges are absent; (iv) All $1 + n + n^2 + n^3$ vertices are isolated.

Figure 2.4: Summary of the results in Theorem 2.13.

Theorem 2.13 (Asymptotic detection of layers in a hierarchical weighted tree). *For each $n \in \mathbb{N}$ let $G_n = (V_n, E_n, w_n)$ be an undirected complete d_n-ary tree with hierarchical edge weights. For each $n \in \mathbb{N}$ let $x_n, y_n \in V_n$ be vertices such that x_n is the parent of y_n and such that the minimal distance between y_n and a leaf of G_n is constant in n. Denote this constant distance by k. Let e_n denote the edge between*

x_n and y_n. For each $n \in \mathbb{N}$ let $q_n > 0$ be the intensity parameter. Then as $n \to \infty$ it holds for the 2-point correlation between x_n and y_n that

$$U_{q_n}^{(G_n)}(x_n, y_n) \to \begin{cases} 0 & \text{if } q_n = o\left(d_n^{-k} w_n(e_n)\right) \\ 1 & \text{if } q_n = \omega\left(d_n^{-k} w_n(e_n)\right). \end{cases} \tag{2.32}$$

2.3 Proofs

2.3.1 Proofs of results on general graphs

2.3.1.1 Monotone events in terms of number of roots

Proof of Lemma 2.2. Let L be the negative graph Laplacian of G. Write $n = |V|$ and let $\lambda_1, \dots, \lambda_n$ denote the eigenvalues $-L$. By [6, proposition 2.1] it holds that

$$\mathbb{E}[r_q] = \sum_{i=1}^{n} \frac{q}{q + \lambda_i} = \frac{qZ'(q)}{Z(q)}, \tag{2.33}$$

so that the derivative of the partition function is given by

$$Z'(q) = \tfrac{1}{q}\mathbb{E}[r_q]Z(q). \tag{2.34}$$

Note that the conditional probability $\mathbb{P}(\Phi_q \in \mathcal{H} \mid r_q = k)$ does not depend on q. Also, the probability $\mathbb{P}(r_q = k)$ can be written as $\frac{c_k q^k}{Z(q)}$, where c_k is some constant independent of q, corresponding to the coefficent of degree k of the characteristic polynomial in (2.4). Hence, we have that

$$\begin{aligned}
\frac{d}{dq}\mathbb{P}(\Phi_q \in \mathcal{H}) &= \frac{d}{dq}\sum_{k=1}^{n} \mathbb{P}(\Phi_q \in \mathcal{H} \mid r_q = k)\mathbb{P}(r_q = k) \\
&= \sum_{k=1}^{n} \mathbb{P}(\Phi_q \in \mathcal{H} \mid r_q = k) c_k \frac{d}{dq}\frac{q^k}{Z(q)} \\
&= \sum_{k=1}^{n} \mathbb{P}(\Phi_q \in \mathcal{H} \mid r_q = k) c_k \frac{kZ(q)q^{k-1} - q^k Z'(q)}{Z(q)^2} \\
&= \tfrac{1}{q}\sum_{k=1}^{n} \mathbb{P}(\Phi_q \in \mathcal{H} \mid r_q = k) c_k \frac{kq^k - q^k \mathbb{E}[r_q]}{Z(q)} \\
&= \tfrac{1}{q}\sum_{k=1}^{n} \mathbb{P}(\Phi_q \in \mathcal{H} \mid r_q = k)\mathbb{P}(r_q = k)\left(k - \mathbb{E}[r_q]\right) \\
&= \tfrac{1}{q}\mathbb{P}(\Phi_q \in \mathcal{H}) \sum_{k=1}^{n} \mathbb{P}(r_q = k \mid \Phi_q \in \mathcal{H})\left(k - \mathbb{E}[r_q]\right) \\
&= \tfrac{1}{q}\mathbb{P}(\Phi_q \in \mathcal{H})\left(\mathbb{E}[r_q \mid \Phi_q \in \mathcal{H}] - \mathbb{E}[r_q]\right),
\end{aligned}$$

where in the last step we use that $\sum_{k=1}^{n} \mathbb{P}(r_q = k \mid \Phi_q \in \mathcal{H}) = 1$. □

2.3.1.2 Monotonicities on undirected networks: proof of Theorem 2.3

Lemma 2.14. *Let $G = (V, E, w)$ be a weighted symmetric graph and let $B \subseteq E$ be a set of directed edges. Then for all $q > 0$ it holds that*

$$\mathbb{E}[r_q \mid \pm B \cap \Phi_q = \varnothing] \geq \mathbb{E}[r_q].$$

Proof. Let $H = G - B$ denote the subgraph of G obtained by removing all edges in B. Let $L^{(G)}$ and $L^{(H)}$ denote the negative graph Laplacians of G and H, respectively. Since these Laplacians are symmetric, $-L^{(G)}$ and $-L^{(H)}$ have real eigenvalues $\lambda_n \geq \ldots \geq \lambda_1$ and $\mu_n \geq \ldots \geq \mu_1$, respectively. By Weyl's monotonicity principle, these eigenvalues satisfy $\lambda_i \geq \mu_i$ for all $i \in [n]$. It follows that $\mathrm{Tr}((qI - L^{(G)})^{-1}) \leq \mathrm{Tr}((qI - L^{(H)})^{-1})$. By the spatial Markov property and [6, prop 2.1] it then holds that

$$\mathbb{E}^{(G)}[r_q \mid B \cap \Phi_q = \varnothing] = \mathbb{E}^{(H)}[r_q] = q\mathrm{Tr}((qI - L^{(H)})^{-1}) \geq q\mathrm{Tr}((qI - L^{(G)})^{-1}) = \mathbb{E}^{(G)}[r_q].$$

□

Lemma 2.15. *Let $G = (V, E, w)$ be a be a weighted symmetric graph and $A \subseteq E$. If $\mathbb{P}(\pm A \subseteq \Phi_q) > 0$, then for all $q > 0$ it holds that*

$$\mathbb{E}[r_q \mid \pm A \subseteq \Phi_q] \leq \mathbb{E}[r_q].$$

Proof. Let $e = (x, y) \in A$ be given. By Lemma A.2 and Lemma 2.14 it holds that

$$\begin{aligned}
&\mathbb{E}[r_q \mid \pm A \subseteq \Phi_q] \\
&= \frac{\mathbb{E}[r_q \mid \pm(A - e) \subseteq \Phi_q] - \mathbb{E}[r_q \mid \pm(A - e) \subseteq \Phi_q, \pm e \notin \Phi_q]\mathbb{P}(\pm e \notin \Phi_q \mid \pm(A - e) \subseteq \Phi_q)}{\mathbb{P}(\pm e \in \Phi_q \mid \pm(A - e) \subseteq \Phi_q)} \\
&= \frac{\mathbb{E}[r_q \mid \pm(A - e) \subseteq \Phi_q] - \mathbb{E}^{(G/(A-e))}[r_q \mid \pm e \notin \Phi_q]\mathbb{P}(\pm e \notin \Phi_q \mid \pm(A - e) \subseteq \Phi_q)}{\mathbb{P}(\pm e \in \Phi_q \mid \pm(A - e) \subseteq \Phi_q)} \\
&\leq \frac{\mathbb{E}[r_q \mid \pm(A - e) \subseteq \Phi_q] - \mathbb{E}^{(G/(A-e))}[r_q]\mathbb{P}(\pm e \notin \Phi_q \mid \pm(A - e) \subseteq \Phi_q)}{\mathbb{P}(\pm e \in \Phi_q \mid \pm(A - e) \subseteq \Phi_q)} \\
&= \frac{\mathbb{E}[r_q \mid \pm(A - e) \subseteq \Phi_q] - \mathbb{E}[r_q \mid \pm(A - e) \subseteq \Phi_q]\mathbb{P}(\pm e \notin \Phi_q \mid \pm(A - e) \subseteq \Phi_q)}{\mathbb{P}(\pm e \in \Phi_q \mid \pm(A - e) \subseteq \Phi_q)} \\
&= \mathbb{E}[r_q \mid \pm(A - e) \subseteq \Phi_q].
\end{aligned}$$

Hence, the result follows by induction on $|A|$. □

Proof of Theorem 2.3. The proof of (2.9) follows directly from Lemmas 2.2 and 2.15.

For the statement about the pairwise connectivity function in (2.10) we argue as follows. Fix $q > 0$. By Lemma 2.2 it is sufficient to show that $\mathbb{E}[r_q] \geq \mathbb{E}[r_q \mid B_q(x) = B_q(y)]$.

Let $\mathcal{P}$ denote the set of undirected paths from x to y, where we interpret a path as a set of directed edges. Then the event $\{B_q(x) = B_q(y)\}$ can be written as the disjoint union

$$\{B_q(x) = B_q(y)\} = \bigcup_{\pi \in \mathcal{P}} \{\pm\pi \subseteq \Phi_q\}.$$

It follows by Lemma 2.15 that

$$\begin{aligned}\mathbb{E}[r_q \mid B_q(x) = B_q(y)] &= \sum_{\pi\in\mathcal{P}} \mathbb{E}[r_q \mid \pm\pi \subseteq \Phi_q]\mathbb{P}(\pm\pi \subseteq \Phi_q \mid B_q(x) = B_q(y)) \\ &\leq \sum_{\pi\in\mathcal{P}} \mathbb{E}[r_q]\mathbb{P}(\pm\pi \subseteq \Phi_q \mid B_q(x) = B_q(y)) = \mathbb{E}[r_q].\end{aligned}$$

□

2.3.2 Two-point correlations on trees

2.3.2.1 Monotonicity of correlations on general trees: proof of Theorem 2.5

Below we show the monotonicity of the 2-point correlation restricted to arbitrary trees. We will start by expressing the 2-point correlation via hitting times in Lemma 2.16. Then in Lemma A.6 we show the monotocity of one point rooting events, by means of Lemma 2.2. After a last bound on the derivatives of hitting time events, given by Lemma A.7, we derive the main claim using these three lemmas.

Lemma 2.16 (Hitting time expression for 2-point correlation between adjacent vertices in trees). *Let $G = (V, E, w)$ be a weighted directed tree and $x, y \in V$ two adjacent vertices. Let $\mathbb{P}_v$ denote the law of the random walk X starting at vertex $v \in V$. The hitting time of vertex v by X is denoted by τ_v and τ_q is an independent exponential killing time with rate q. Then it holds that*

$$U_q^{(G)}(x, y) = \frac{1 - \mathbb{P}_x(\tau_y < \tau_q) - \mathbb{P}_y(\tau_x < \tau_q) + \mathbb{P}_x(\tau_y < \tau_q)\mathbb{P}_y(\tau_x < \tau_q)}{1 - \mathbb{P}_x(\tau_y < \tau_q)\mathbb{P}_y(\tau_x < \tau_q)}.$$

Proof of Lemma 2.16. We will reason using the representation in (2.8) coming from Wilson's sampling construction. We note in particular that in order for the directed edge (x, y) to be present in Φ_q, it is equivalent to require that the loop-erased trajectory in (2.8) includes y, which can be expressed in terms of hitting times of the random walk as

$$\begin{aligned}\mathbb{P}((x, y) \in \Phi_q) &= \mathbb{P}_x(\tau_y < \tau_q) \sum_{k=0}^{\infty} \left(\mathbb{P}_y(\tau_x < \tau_q)\mathbb{P}_x(\tau_y < \tau_q)\right)^k \mathbb{P}_y(\tau_q < \tau_x) \\ &= \frac{\mathbb{P}_x(\tau_y < \tau_q)\left(1 - \mathbb{P}_y(\tau_x < \tau_q)\right)}{1 - \mathbb{P}_x(\tau_y < \tau_q)\mathbb{P}_y(\tau_x < \tau_q)},\end{aligned}$$

where the index k in the above sum represents the number of times that the random walk reaches y and then does return to x. We notice in particular that the above step is equivalent to use the forest transfer-current kernel in [5]. For the reversed edge (y, x), we can write

$$\mathbb{P}((y, x) \in \Phi_q) = \left(1 - \mathbb{P}((x, y) \in \Phi_q)\right)\mathbb{P}_y(\tau_x < \tau_q),$$

where these two factors correspond to (2.8). Therefore, it follows that

$$\begin{aligned} U_q^{(G)}(x,y) &= 1 - \mathbb{P}((x,y) \in \Phi_q) - \mathbb{P}((y,x) \in \Phi_q) \\ &= 1 - \tfrac{\mathbb{P}_x(\tau_y<\tau_q)(1-\mathbb{P}_y(\tau_x<\tau_q))}{1-\mathbb{P}_x(\tau_y<\tau_q)\mathbb{P}_y(\tau_x<\tau_q)} - \left(1 - \tfrac{\mathbb{P}_x(\tau_y<\tau_q)(1-\mathbb{P}_y(\tau_x<\tau_q))}{1-\mathbb{P}_x(\tau_y<\tau_q)\mathbb{P}_y(\tau_x<\tau_q)}\right) \mathbb{P}_y(\tau_x < \tau_q) \\ &= \frac{1 - \mathbb{P}_x(\tau_y < \tau_q) - \mathbb{P}_y(\tau_x < \tau_q) + \mathbb{P}_x(\tau_y < \tau_q)\mathbb{P}_y(\tau_x < \tau_q)}{1 - \mathbb{P}_x(\tau_y < \tau_q)\mathbb{P}_y(\tau_x < \tau_q)}. \end{aligned}$$

□

For brevity the proofs below use the notation $\{x \leftrightarrow y\} := \{B_q(x) = B_q(y)\}$ to denote the event that x and y are connected in Φ_q. We write $\{x \nleftrightarrow y\}$ to denote the complementary event.

Proof of Theorem 2.5. Let $d = d(x,y)$ denote the distance between x and y in G and let z be the vertex adjacent to x with distance $d-1$ to y. We proceed by induction on d.

If $d = 1$, then by Lemma 2.16 we have that

$$U_q(x,y) = \frac{1 - \mathbb{P}_x(\tau_y < \tau_q) - \mathbb{P}_y(\tau_x < \tau_q) + \mathbb{P}_x(\tau_y < \tau_q)\mathbb{P}_y(\tau_x < \tau_q)}{1 - \mathbb{P}_x(\tau_y < \tau_q)\mathbb{P}_y(\tau_x < \tau_q)}.$$

Taking the derivative gives us that

$$\frac{d}{dq}U_q(x,y) = -\frac{(1 - \mathbb{P}_x(\tau_y < \tau_q))^2 \frac{d}{dq}\mathbb{P}_y(\tau_x < \tau_q) + (1 - \mathbb{P}_y(\tau_x < \tau_q))^2 \frac{d}{dq}\mathbb{P}_x(\tau_y < \tau_q)}{\left(1 + \mathbb{P}_x(\tau_y < \tau_q)\mathbb{P}_y(\tau_x < \tau_q)\right)^2},$$

which is non-negative by the upper bound in Lemma A.5.

Now assume that $d \geq 2$. We then have that

$$\begin{aligned} &\frac{d}{dq}U_q(x,y) \\ &= \frac{d}{dq}\left(\mathbb{P}(x \nleftrightarrow z \mid z \leftrightarrow y)\mathbb{P}(z \leftrightarrow y) + \mathbb{P}(z \nleftrightarrow y)\right) \\ &= \mathbb{P}(x \nleftrightarrow z \mid z \leftrightarrow y)\tfrac{d}{dq}\mathbb{P}(z \leftrightarrow y) + \mathbb{P}(z \leftrightarrow y)\tfrac{d}{dq}\mathbb{P}(x \nleftrightarrow z \mid z \leftrightarrow y) + \tfrac{d}{dq}\mathbb{P}(z \nleftrightarrow y) \\ &= (1 - \mathbb{P}(x \nleftrightarrow z \mid z \leftrightarrow y))\tfrac{d}{dq}\mathbb{P}(z \nleftrightarrow y) + \mathbb{P}(z \leftrightarrow y)\tfrac{d}{dq}\mathbb{P}(x \nleftrightarrow z \mid z \leftrightarrow y). \end{aligned}$$

By the induction hypothesis, we have that $\frac{d}{dq}\mathbb{P}(z \nleftrightarrow y) \geq 0$. Hence, it remains to show that

$$\tfrac{d}{dq}\mathbb{P}(x \nleftrightarrow z \mid z \leftrightarrow y) \geq 0.$$

Removing the edges between x and z splits G into two connected components. Let T_x and T_z denote the components containing vertex x and z, respectively. By Lemma A.4 it then holds that

$$\mathbb{P}(x \nleftrightarrow z \mid z \leftrightarrow y) = \frac{q}{q + w(x,z)\mathbb{P}^{(T_x)}(x \in R_q) + w(z,y)\mathbb{P}^{(T_z)}(z \in R_q \mid z \leftrightarrow y)}.$$

Taking the derivative and applying Lemmas A.6 and A.7 gives us that

$$\begin{aligned}\frac{d}{dq}\mathbb{P}(x \nleftrightarrow z \mid z \leftrightarrow y) &= \frac{w(x,z)\mathbb{P}^{(T_x)}(x \in R_q) + w(z,y)\mathbb{P}^{(T_z)}(z \in R_q \mid z \leftrightarrow y)}{\big(q + w(x,z)\mathbb{P}^{(T_x)}(x \in R_q) + w(z,y)\mathbb{P}^{(T_z)}(z \in R_q \mid z \leftrightarrow y)\big)^2} \\ &\quad - \frac{qw(x,z)\frac{d}{dq}\mathbb{P}^{(T_x)}(x \in R_q) + qw(z,y)\frac{d}{dq}\mathbb{P}^{(T_z)}(z \in R_q \mid z \leftrightarrow y)}{\big(q + w(x,z)\mathbb{P}^{(T_x)}(x \in R_q) + w(z,y)\mathbb{P}^{(T_z)}(z \in R_q \mid z \leftrightarrow y)\big)^2} \\ &\quad \geq 0.\end{aligned}$$

□

2.3.2.2 Inclusion-exclusion for connectivity function on general trees

Proof of Theorem 2.6. We will prove the statement by induction on d. First assume that $d = 1$. Write $\mathcal{H} = \{F \in \mathcal{F}_G \colon\ x \nleftrightarrow_F y\}$ to denote the set of rooted forests not containing an edge between x and y. Since G is a tree, removing the edges between x and y yields two connected components G_x and G_y containing vertex x and vertex y, respectively. Note that for the non-normalized measure on G it holds that $\nu^{(G)}(\Phi_q = F) = \nu_q^{(G_x)}(\Phi = F[G_x])\nu_q^{(G_y)}(\Phi = F[G_y])$ for all $F \in \mathcal{H}$, where $F[G_x]$ and $F[G_x]$ denote the induced subgraphs of F on the vertices of G_x and G_y, respectively. For all $F_x \in \mathcal{F}_{G_x}$ and $F_y \in \mathcal{F}_{G_y}$ there is exactly one $F \in \mathcal{H}$ with $F[G_x] = F_x$ and $F[G_y] = F_y$, namely the disjoint graph union of F_x and F_y. Hence, it holds that

$$\begin{aligned}U_q^{(G)}(x,y) &= \sum_{F \in \mathcal{H}} \mathbb{P}^{(G)}(\Phi_q = F) = \frac{1}{Z_G(q)} \sum_{F \in \mathcal{H}} \nu^{(G_x)}(\Phi_q = F[G_x])\nu^{(G_y)}(\Phi_q = F[G_y]) \\ &= \frac{1}{Z_G(q)} \sum_{F_x \in \mathcal{F}_{G_x}} \sum_{F_y \in \mathcal{F}_{G_y}} \nu^{(G_x)}(\Phi_q = F_x)\nu^{(G_y)}(\Phi_q = F_y) = \frac{Z_{G_x}(q)Z_{G_y}(q)}{Z_G(q)}.\end{aligned}$$

Now assume that $d > 1$. Let $z = z_{d-1}$ denote the neighbor of y with distance $d-1$ to x. Let $G_{\{d\}}$ denote the graph obtained from G by removing the edges between y and z. Then $G_{\{d\}}$ consists of two components G_y and G_z containing vertex y and z

respectively. It then holds by Lemma A.2 and the induction hypothesis that

$$\begin{aligned}
U_q^{(G)}(x,y) &= U_q^{(G)}(x,z) + U_q^{(G)}(y,z) - \mathbb{P}^{(G)}(x \nleftrightarrow_{\Phi_q} z,\ y \nleftrightarrow_{\Phi_q} z) \\
&= U_q^{(G)}(x,z) + U_q^{(G)}(y,z) - U_q^{(G)}(y,z)\mathbb{P}^{(G)}(x \nleftrightarrow_{\Phi_q} z \mid y \nleftrightarrow_{\Phi_q} z) \\
&= U_q^{(G)}(x,z) + U_q^{(G)}(y,z) - U_q^{(G)}(y,z)U_q^{(G_{\{d\}})}(x,z) \\
&= \frac{1}{Z_G(q)}\left(\sum_{k=1}^{d-1}(-1)^{k+1}\sum_{I\in\binom{[d-1]}{k}}\prod_{i=1}^{k+1} Z_{G_I^i}(q)\right) + \frac{Z_{G_y}(q)Z_{G_z}(q)}{Z_G(q)} \\
&\quad - \frac{Z_{G_y}(q)Z_{G_z}(q)}{Z_G(q)}\frac{1}{Z_{G_{\{d\}}}(q)}\left(\sum_{k=1}^{d-1}(-1)^{k+1}\sum_{I\in\binom{[d-1]}{k}}\prod_{i=1}^{k+2} Z_{G^i_{I\cup\{d\}}}(q)\right) \\
&= \frac{1}{Z_G(q)}\left(\sum_{k=1}^{d-1}(-1)^{k+1}\sum_{I\in\binom{[d-1]}{k}}\prod_{i=1}^{k+1} Z_{G_I^i}(q)\right) + \frac{Z_{G_y}(q)Z_{G_z}(q)}{Z_G(q)} \\
&\quad + \frac{1}{Z_G(q)}\left(\sum_{k=1}^{d-1}(-1)^{k}\sum_{I\in\binom{[d-1]}{k}}\prod_{i=1}^{k+2} Z_{G^i_{I\cup\{d\}}}(q)\right) \\
&= \frac{1}{Z_G(q)}\left(\sum_{k=1}^{d-1}(-1)^{k+1}\sum_{I\in\binom{[d-1]}{k}}\prod_{i=1}^{k+1} Z_{G_I^i}(q)\right) \\
&\quad + \frac{1}{Z_G(q)}\left(\sum_{k=0}^{d-1}(-1)^{k}\sum_{I\in\binom{[d-1]}{k}}\prod_{i=1}^{k+2} Z_{G^i_{I\cup\{d\}}}(q)\right) \\
&= \frac{1}{Z_G(q)}\left(\sum_{k=1}^{d}(-1)^{k+1}\sum_{I\in\binom{[d]}{k}}\prod_{i=1}^{k+1} Z_{G_I^i}(q)\right).
\end{aligned}$$

□

2.3.3 Partition function on segments and rings

Proof of Theorem 2.7.

Equation (2.13) Let b be a boundary vertex of PG_n. Let $\nu^{(n)}$ and Z_n denote the non-normalized measure and partition function of PG_n, respectively. By Lemma A.4 we have that

$$\nu^{(n)}(b \notin R_q) = Z_{n-1}(q), \text{ and } \nu^{(n)}(b \in R_q) = \nu^{(n-1)}(b \in R_q) + qZ_{n-1}(q). \tag{2.35}$$

This gives us that

$$\begin{aligned}
Z_n(q) &= \nu^{(n)}(b \in R_q) + \nu^{(n)}(b \notin R_q) = \nu^{(n-1)}(b \in R_q) + (q+1)Z_{n-1}(q) \\
&= (q+2)Z_{n-1}(q) - \nu^{(n-1)}(b \notin R_q) = (q+2)Z_{n-1}(q) - Z_{n-2}(q).
\end{aligned} \tag{2.36}$$

We will prove Equation (2.13) by induction on n. Note that for $n = 1$ we have $Z_1(q) = q$ and for $n = 2$ we have $Z_2(q) = q^2+2q$, so in both these cases Equation (2.13) holds. Now assume that $n > 2$. Then by Equation (2.36), the induction hypothesis and repeated applications of Pascal's formula we have that

$$Z_n(q) = (2+q)Z_{n-1}(q) - Z_{n-2}(q) = (2+q)\sum_{k=1}^{n-1}\binom{n+k-2}{2k-1}q^k - \sum_{k=1}^{n-2}\binom{n+k-3}{2k-1}q^k$$
$$= \sum_{k=1}^{n}\binom{n+k-1}{2k-1}q^k.$$

Equation (2.14) Let L denote the negative graph Laplacian of PG_n, since due to (2.4) the partition function is the characteristic polynomial of L, it can be directly obtained from its spectrum, which is given in [64], from which:

$$Z_n(q) = \prod_{k=1}^{n}\left(q + 2 - 2\cos\left(\tfrac{\pi(n-k)}{n}\right)\right).$$

Equation (2.15) We have shown above that the partition function satisfies the recurrence relation in Equation (2.36) . Using the initial conditions $Z_1(q) = q$ and $Z_2(q) = q^2 + 2q$, this linear recurrence relation has solution

$$Z_n(q) = \frac{q\left(q+2+\sqrt{q^2+4q}\right)^n - q\left(q+2-\sqrt{q^2+4q}\right)^n}{2^n\sqrt{q^2+4q}}.$$

Equation (2.16) To verify that the three expressions above do indeed coincide, we can use Chebyshev polynomials of the second kind and find that

$$Z_n(q) = qT_{n-1}(\tfrac{q}{2}+1).$$

□

We next move to the proof of Corollary 2.7.1, for which we will first need to expresses in the next lemma the probability of a boundary point in the path-graph being a root in terms of differences of the partition function.

Lemma 2.17 (Rooting events in path-graphs). *Let PG_n be the path-graph on n vertices and $Z_n(q)$ its partition function. Let $x \in V$ be a vertex with distance $d \in \mathbb{N}_0$ from the boundary and $b \in V$ a boundary vertex. Let $\nu^{(n)}$ denote the non-normalized measure on PG_n and R_q the set of roots of Φ_q. Then*

$$\nu^{(n)}(x \in R_q) = \tfrac{1}{q}\nu^{(d+1)}(b \in R_q)\ \nu^{(n-d)}(b \in R_q), \tag{2.37}$$

with

$$\nu^{(n)}(b \in R_q) = Z_n(q) - Z_{n-1}(q). \tag{2.38}$$

For the non-normalized measure of the event that both boundary vertices b and b' are roots it holds that

$$\nu^{(n)}(b, b' \in R_q) = qZ_{n-1}(q). \tag{2.39}$$

Proof of Lemma 2.17.

Equation (2.37) Let L_n denote the negative graph Laplacian of the path-graph on n vertices. Inspection of the Laplacian and using the symmetry of the path-graph shows that

$$\det[qI - L_n]_x = \det[qI - L_{d+1}]_b \det[qI - L_{n-d}]_b,$$

as removing a row and column from $qI - L_n$ results in a matrix comprised of two blocks. Since the event that vertex x is a root equals the event that none of the outgoing edges of x are present, it holds by Lemma A.2 that $\nu^{(n)}(x \in R_q) = q \ \det[qI - L_n]_x$, from which Equation (2.37) follows.

Equation (2.38) Since $\nu^{(n)}(b \in R_q) = Z_n(q) - \nu^{(n)}(b \notin R_q)$, Equation (2.38) follows directly from Equation (2.35).

Equation (2.39) By Lemma A.4 we have that

$$\begin{aligned}\nu^{(n)}(b, b' \in R_q) &= q\nu^{(n-1)}(b \in R_q) + \nu^{(n-1)}(b, b' \in R_q),\\ \nu^{(n)}(b \in R_q,\ b' \notin R_q) &= \nu^{(n-1)}(b \in R_q).\end{aligned}$$

Since $\nu^{(n-1)}(b, b' \in R_q) = \nu^{(n-1)}(b \in R_q) - \nu^{(n-1)}(b \in R_q,\ b' \notin R_q)$, it follows from Equations (2.36) and (2.38) that

$$\begin{aligned}\nu^{(n)}(b, b' \in R_q) &= (q+1)\nu^{(n-1)}(b \in R_q) - \nu^{(n-2)}(b \in R_q)\\ &= (q+1)Z_{n-1}(q) - (q+2)Z_{n-2}(q) + Z_{n-3}(q) = qZ_{n-1}(q).\end{aligned}$$

□

Proof of Corollary 2.7.1. We will first prove Equation (2.17). Let V denote the vertex set of CG_n and let $x \in V$ be a vertex. The partition function can be split into two terms

$$Z_{CG_n}(q) = \nu^{(CG_n)}(x \in R_q) + \nu^{(CG_n)}(x \notin R_q). \tag{2.40}$$

Note that the induced subgraph $CG_n[V \setminus \{x\}]$ obtained by removing vertex x, is a path-graph on $n-1$ vertices. Let y and z denote the two vertices adjacent to x in CG_n. So, these are the boundary vertices of PG_{n-1}. We will use Lemma A.3. This gives us by Equation (2.36) and lemma 2.17 that

$$\begin{aligned}\nu^{(CG_n)}(x \in R_q) &= \sum_{F \in \mathcal{F}_{PG_{n-1}}} q\ \nu^{(PG_{n-1})}(\Phi = F)\ (1 + \tfrac{1}{q})^{|R(F) \cap \{y,z\}|}\\ &= (q + 2 + \tfrac{1}{q})\nu^{(PG_{n-1})}(y, z \in R_q) + 2(q+1)\nu^{(PG_{n-1})}(y \in R_q, z \notin R_q)\\ &\quad + q\nu^{(PG_{n-1})}(y, z \notin R_q)\\ &= qZ_{PG_{n-1}}(q) + (2 + \tfrac{1}{q})\nu^{(PG_{n-1})}(y, z \in R_q) + 2\nu^{(PG_{n-1})}(y \in R_q, z \notin R_q)\\ &= (q+2)Z_{PG_{n-1}}(q) - 2Z_{PG_{n-2}}(q) + \tfrac{1}{q}\nu^{(PG_{n-1})}(y, z \in R_q)\\ &= Z_{PG_n}(q) - Z_{PG_{n-2}}(q) + \tfrac{1}{q}\nu^{(PG_{n-1})}(y, z \in R_q) = Z_{PG_n}(q).\end{aligned}$$

Let $r_y(F)$ denote the root in the tree of forest F that contains vertex y. Again using Lemma A.3 and Equation (2.36), we obtain

$$\begin{aligned}
\nu^{(CG_n)}(x \notin R_q) &= \sum_{F \in \mathcal{F}_{PG_{n-1}}} \nu^{(PG_{n-1})}(\Phi = F)\ 2(1+\tfrac{1}{q})^{\mathbf{1}\{z \in R(F),\ r_y(F) \neq z\}} \\
&= 2\nu^{(PG_{n-1})}(z \notin R_q) + 2(1+\tfrac{1}{q})\nu^{(PG_{n-1})}(z \in R_q, r_y(F) \neq z) \\
&= (2+\tfrac{2}{q})Z_{PG_{n-1}}(q) - \tfrac{2}{q}Z_{PG_{n-2}}(q) - 2 \\
&= \tfrac{2}{q}((q+2)Z_{PG_{n-1}}(q) - Z_{PG_{n-2}}(q) - Z_{PG_{n-1}}(q)) - 2 \\
&= \tfrac{2}{q}(Z_{PG_n}(q) - Z_{PG_{n-1}}(q)) - 2.
\end{aligned}$$

This proves Equation (2.17).

Equation (2.18) follows from Equation (2.17) and the expression for the path-graph partition function given in Equation (2.13), by repeated applications of Pascal's formula. □

2.3.3.1 Asymptotic analysis of path-graphs

Proof of Theorem 2.8.

Equation (2.19) Let $\mathcal{F}_n$ denote the set of rooted forests of PG_n and write

$$\begin{aligned}
\mathcal{F}^k_{n-d} &= \{F \in \mathcal{F}_{n-d} \colon\ r(F) = k\}; \\
\mathcal{R}^k_{n-d}(x) &= \{F \in \mathcal{F}_{n-d} \colon\ r(F) = k,\ x \in R(F)\}; \\
\mathcal{C}^k_n(x,y) &= \{F \in \mathcal{F}_n \colon\ r(F) = k,\ x \leftrightarrow_F y\}.
\end{aligned}$$

It is sufficient to show that for all $k \in [n-d]$ it holds that $|\mathcal{C}^k_n(x,y)| = |\mathcal{F}^k_{n-d}| + d|\mathcal{R}^k_{n-d}(x)|$. The result then follows from Lemma 2.17.

We will construct a bijection between the set $\mathcal{C}^k_n(x,y)$ and the set $\mathcal{F}^k_{n-d} \cup (\mathcal{R}^k_{n-d}(x) \times [d])$. Let $F \in \mathcal{C}^k_n(x,y)$ be given. Let $r \in [n]$ denote the vertex of F that is the root in the component of x and y. Let $B = [y] \setminus [x-1]$ denote the set of all vertices from x to y and let $F_B \in \mathcal{F}^k_{n-d}$ denote the B-vertex contraction of F. Then we have that $F_B \in \mathcal{R}^k_{n-d}(x)$ if and only if $r \in [y] \setminus [x-1]$. Define the function $f : \mathcal{C}^k_n(x,y) \to \mathcal{F}^k_{n-d} \cup (\mathcal{R}^k_{n-d}(x) \times [d])$ by

$$f(F) = \begin{cases} F_B & \text{if } r \notin [y] \setminus [x] \\ (F_B, r-x) & \text{if } r \in [y] \setminus [x]. \end{cases}$$

It is easily verified that this gives a bijection.

Lower bound Let $\tilde{\mathbb{P}}_x$ denote the law of the discrete-time random walk $\tilde{X}$ on PG_n starting on x, as defined in Equation (A.6). Since in this case we consider a path-graph, we have that $\tilde{\tau}_q \sim \mathrm{Geom}(\frac{q}{q+2})$.

We will analyze the expression in Equation (2.8). Let z denote a vertex halfway between x and y. For notational simplicity we assume that d is even, so that $z = x + \frac{d}{2}$. The argument in the case where d is odd is similar. Note that the vertices x and y

are disconnected in Φ_q if both the random walks starting at x and the random walk starting at y are killed before reaching vertex z. So, we have that

$$\mathbb{P}(x \nleftrightarrow_{\Phi_q} y) \geq \tilde{\mathbb{P}}_x(\tilde{\tau}_q \leq \tau_z)\tilde{\mathbb{P}}_y(\tilde{\tau}_q \leq \tau_z). \tag{2.41}$$

Let $\tau_S(k)$ denote the hitting time of $k \in \mathbb{Z}$ by S. A coupling of $\tilde{X}$ and S can be used to show that

$$\tau_z \stackrel{d}{=} \min\{\tau_S\left(\tfrac{d}{2}\right), \tau_S\left(1 - 2x - \tfrac{d}{2}\right)\}, \tag{2.42}$$

where $\stackrel{d}{=}$ denotes equality in distribution.

By the reflection principle it holds for all $k, n \in \mathbb{N}$ that

$$\mathbb{P}(\tau_S(n) \leq k) = \mathbb{P}(S_k \notin [-n, n-1]).$$

For $u \in \{x, y\}$ it follows that

$$\begin{aligned}
&\tilde{\mathbb{P}}_u(\tilde{\tau}_q \leq \tau_z) = \sum_{k=1}^{\infty} \mathbb{P}(\tilde{\tau}_q = k)\tilde{\mathbb{P}}_u(\tau_z \geq k) \geq \sum_{k=1}^{m} \left(1 - \tilde{\mathbb{P}}_u(\tau_z < k)\right) \mathbb{P}(\tilde{\tau}_q = k) \\
&= \sum_{k=1}^{m} \left(1 - \mathbb{P}(\tau_S(\tfrac{d}{2}) < k \text{ or } \tau_S(1 - 2x - \tfrac{d}{2}) < k)\right) \mathbb{P}(\tilde{\tau}_q = k) \\
&\geq \sum_{k=1}^{m} \left(1 - 2\mathbb{P}(\tau_S(\tfrac{d}{2}) < k)\right) \mathbb{P}(\tilde{\tau}_q = k) = \sum_{k=1}^{m} \left(2\mathbb{P}(S_{k-1} \in [-\tfrac{d}{2}, \tfrac{d}{2} - 1]) - 1\right) \mathbb{P}(\tilde{\tau}_q = k) \\
&\geq \sum_{k=1}^{m} \left(2\mathbb{P}(|S_{k-1} < \tfrac{d}{2}) - 1\right) \mathbb{P}(\tilde{\tau}_q = k) \geq \sum_{k=1}^{m} \left(2\mathbb{P}(|S_m| < \tfrac{d}{2}) - 1\right) \mathbb{P}(\tilde{\tau}_q = k) \\
&= \left(2\mathbb{P}\left(|S_m| < \tfrac{d}{2}\right) - 1\right) \mathbb{P}(\tilde{\tau}_q \leq m) = \left(2\mathbb{P}\left(|S_m| < \tfrac{d}{2}\right) - 1\right) \left(1 - \left(1 - \tfrac{q}{2+q}\right)^m\right).
\end{aligned}$$

Hence,

$$\min_{u \in \{x,y\}} \tilde{\mathbb{P}}_u(\tilde{\tau}_q \leq \tau_z) \geq \left(2\mathbb{P}\left(|S_m| < \tfrac{d}{2}\right) - 1\right) \left(1 - \left(1 - \tfrac{q}{2+q}\right)^m\right).$$

If $\left(2\mathbb{P}\left(|S_m| < \frac{d}{2}\right) - 1\right)$ is non-negative, then we also have that

$$\tilde{\mathbb{P}}_x(\tilde{\tau}_q \leq \tau_z)\tilde{\mathbb{P}}_y(\tilde{\tau}_q \leq \tau_z) \geq \left(2\mathbb{P}\left(|S_m| < \tfrac{d}{2}\right) - 1\right)^2 \left(1 - \left(1 - \tfrac{q}{2+q}\right)^m\right)^2.$$

Therefore, we have for all $m \in \mathbb{N}$ with $\mathbb{P}\left(|S_m| < \frac{d}{2}\right) \geq \frac{1}{2}$ that

$$U_q^{(n)}(x, y) \geq \left(2\mathbb{P}\left(|S_m| < \tfrac{d}{2}\right) - 1\right)^2 \left(1 - \left(1 - \tfrac{q}{2+q}\right)^m\right)^2,$$

which gives the desired lower bound.

Upper bound We again analyze by means of Wilson's algorithm with the first random walk starting at x and the second one starting at y. Note that the trajectory of the loop-erasure of the first random walk will always contain its starting vertex x. Thus

if the second random walk hits x before being killed, then x and y are connected in Φ_q. Therefore, we have that

$$\mathbb{P}(x \leftrightarrow_{\Phi_q} y) \geq \tilde{\mathbb{P}}_y(\tau_x < \tilde{\tau}_q).$$

Using a coupling argument we can show that

$$\tau_x \stackrel{d}{=} \min\{\tau_S(-d), \tau_S(2n + d - 2y + 1)\},$$

where τ_x denotes the first hitting time vertex x by the random walk $\tilde{X}$ starting at y. So, in a manner similar to that used for the lower bound, we find for all $m \in \mathbb{N}$ that

$$\begin{aligned}
\tilde{\mathbb{P}}_y(\tau_x < \tilde{\tau}_q) &= \sum_{k=1}^{\infty} \tilde{\mathbb{P}}_y(\tau_x < k)\mathbb{P}(\tilde{\tau}_q = k) \geq \sum_{k=m}^{\infty} \tilde{\mathbb{P}}_y(\tau_x \leq k)\mathbb{P}(\tilde{\tau}_q = k+1) \\
&= \sum_{k=m}^{\infty} \mathbb{P}(\tau_S(-d) \leq k \text{ or } \tau_S(2n + d - 2y + 1) \leq k)\mathbb{P}(\tilde{\tau}_q = k+1) \\
&\geq \sum_{k=m}^{\infty} \mathbb{P}(\tau_S(-d) \leq k)\mathbb{P}(\tilde{\tau}_q = k+1) \geq \sum_{k=m}^{\infty} \mathbb{P}(\tau_S(d) \leq m)\mathbb{P}(\tilde{\tau}_q = k+1) \\
&= \mathbb{P}(\tau_S(d) \leq m)\mathbb{P}(\tilde{\tau}_q > m) = \mathbb{P}(S_m \notin [-d, d-1])\mathbb{P}(\tilde{\tau}_q > m) \\
&\geq \mathbb{P}(|S_m| > d)\mathbb{P}(\tilde{\tau}_q > m) = \mathbb{P}\left(|S_m| > d\right)\left(1 - \tfrac{q}{2+q}\right)^m.
\end{aligned}$$

It follows that

$$U_q^{(n)}(x, y) = 1 - \mathbb{P}(x \leftrightarrow_{\Phi_q} y) \leq 1 - \mathbb{P}\left(|S_m| > d\right)\left(1 - \tfrac{q}{2+q}\right)^m.$$

□

Proof of Corollary 2.8.1.

$\boxed{q_n = o(\frac{1}{d_n^2})}$ Set $m_n = \left\lceil \frac{d_n}{\sqrt{q_n}} \right\rceil$, i.e. m_n is the smallest integer that is not smaller than $\frac{d_n}{\sqrt{q_n}}$. We have that $m_n = \omega(d_n^2)$. In particular this means that $m_n \to \infty$ as $n \to \infty$. So, $\frac{S_{m_n}}{\sqrt{m_n}}$ converges in distribution to a standard normal random variable. Since $\frac{d_n}{\sqrt{m_n}} \to 0$, it follows that $\mathbb{P}\left(\frac{|S_{m_n}|}{\sqrt{m_n}} > \frac{d_n}{\sqrt{m_n}}\right) \to 1$. We also have that $m_n = o(\frac{1}{q_n})$, which gives us that $\left(1 - \frac{q_n}{2+q_n}\right)^{m_n} \to 1$. Therefore, the upper bound from Theorem 2.8 gives us that

$$U_{q_n}^{(n)}(x_n, y_n) \leq 1 - \mathbb{P}\left(\tfrac{|S_{m_n}|}{\sqrt{m_n}} > \tfrac{d_n}{\sqrt{m_n}}\right)\left(1 - \frac{q_n}{2 + q_n}\right)^{m_n} = o(1).$$

$\boxed{q_n = \omega(\frac{1}{d_n^2})}$ Again set $m_n = \left\lceil \frac{d_n}{\sqrt{q_n}} \right\rceil$. It holds that $m_n = \omega(\frac{1}{q_n})$, and hence that $\left(1 - \frac{q_n}{2+q_n}\right)^{m_n} \to 0$. Furthermore, we have that $m_n = o(d_n^2)$. This means that $\frac{d_n}{2\sqrt{m_n}} \to \infty$ and thus that $\mathbb{P}\left(\frac{|S_{m_n}|}{\sqrt{m_n}} < \frac{d_n}{2\sqrt{m_n}}\right) \to 1$. For large enough n, this gives

us that $\mathbb{P}\left(\frac{|S_{m_n}|}{\sqrt{m_n}} < \frac{d_n}{2\sqrt{m_n}}\right) \geq \frac{1}{2}$, which means that we can apply the lower bound from Theorem 2.8. This gives us that

$$U_{q_n}^{(n)}(x_n, y_n) \geq \left(1 - \left(1 - \frac{q_n}{2+q_n}\right)^{m_n}\right)^2 \left(2\mathbb{P}\left(\frac{|S_{m_n}|}{\sqrt{m_n}} < \frac{d_n}{2\sqrt{m_n}}\right) - 1\right)^2 = 1 - o(1).$$

$\boxed{q_n = \frac{c}{d_n^2} + o(\frac{1}{d_n^2})}$ Now set $m_n = \left\lceil \frac{d_n}{4\sqrt{cq_n}} \right\rceil$. We will distinguish between the case where d_n diverges and the case where d_n is bounded.

First assume that $d_n = \omega(1)$. Then we find that $m_n \sim \frac{1}{4q_n}$. It follows that there exists an $\varepsilon > 0$ small enough that $\varepsilon < \left(1 - \frac{q_n}{2+q_n}\right)^{m_n} < 1 - \varepsilon$ for all $n \in \mathbb{N}$. We also have that $m_n \sim \frac{1}{4}d_n^2$. This gives us that $\frac{S_{m_n}}{\sqrt{m_n}}$ converges in distribution to a standard normal random variable Z and that $\frac{d_n}{2\sqrt{m_n}} \to 1$. Since $0.6 < \mathbb{P}(|Z| < 1) < 0.7$, we can apply the lower bound from Theorem 2.8. Using both bounds from Theorem 2.8, we conclude the non-degeneracy:

$$\lim_{n\to\infty} U_{q_n}^{(n)}(x_n, y_n) \in (0, 1).$$

Now instead assume that d_n is bounded, i.e. there exists an $M \in \mathbb{N}$ with $M \geq d_n$ for all $n \in \mathbb{N}$. Then the lower bound from Theorem 2.8 can not necessarily be applied. However, we can lower bound the probability that x_n and y_n are disconnected by the probability that the discrete-time random walks on PG_n starting at x and y are both killed at time 1, while still at their starting points. This probability equals $\mathbb{P}(\tilde{\tau}_q = 1)^2 = \frac{q_n^2}{(2+q_n)^2}$.

The probability that x_n and y_n are connected can be lower bounded by the probability that the discrete-time random walk on PG_n starting at x jumps M times in the direction of y and then is then killed at time $M + 1$. This probability equals $\left(\frac{1}{2+q_n}\right)^M \frac{q_n}{2+q_n}$. So, we have for all $n \in \mathbb{N}$ that

$$\frac{q_n^2}{(2+q_n)^2} \leq U_{q_n}^{(n)}(x_n, y_n) \leq 1 - \left(\frac{1}{2+q_n}\right)^M \frac{q_n}{2+q_n}.$$

Since $q_n \sim \frac{c}{d_n^2}$ and d_n is bounded, we have that q_n is bounded away from 0 and away from infinity. Hence, the 2-point correlation is also non-degenerate in this case. □

Proof of Equation (2.21). For brevity write $Z_n(q) = Z_{PG_n}(q)$ and set $a_n = \sqrt{q_n^2 + 4q_n}$. Using the expression for the partition function given in Equation (2.15) we have for each $m \in \mathbb{N}$ that

$$Z_m(q_n) = \frac{1}{\sqrt{1+4d_n^2}} \left(1 - \left(\frac{q_n + 2 - a_n}{q_n + 2 + a_n}\right)^m\right) \left(\frac{q_n + 2 + a_n}{2}\right)^m. \tag{2.43}$$

By Equation (2.19) the 2-point correlation is given by

$$U_{q_n}^{(G_n)}(x_n, y_n) = 1 - \frac{Z_{n-d_n}(q_n)}{Z_n(q_n)} - \frac{d_n \left[Z_{x_n}(q_n) - Z_{x_n-1}(q_n)\right] \left[Z_{n-y_n+1}(q_n) - Z_{n-y_n}(q_n)\right]}{q_n Z_n(q_n)}. \tag{2.44}$$

The result follows by plugging in Equation (2.43) into Equation (2.44) and repeatedly applying the following limits:

$$\lim_{n\to\infty}\left(\frac{2}{q_n+2+a_n}\right)^{\alpha\sqrt{n}+o(\sqrt{n})} = e^{-\frac{\alpha}{2\delta}}, \qquad \lim_{n\to\infty}\left(\frac{q_n+2-a_n}{q_n+2+a_n}\right)^{\alpha\sqrt{n}+o(\sqrt{n})} = e^{-\frac{\alpha}{\delta}},$$

$$\text{and} \qquad \lim_{n\to\infty}\left(\frac{q_n+2-a_n}{q_n+2+a_n}\right)^{\omega(\sqrt{n})} = 0,$$

with $\alpha \in \mathbb{R}$ being a constant. To conclude, let us check the correctness of these last three simple limits.

For the first one, since $\frac{\sqrt{1+4d_n^2}}{2d_n^2} - \frac{1}{2\delta\sqrt{n}} = o(\frac{1}{\sqrt{n}})$, we have that

$$\left(\frac{q_n+2+a_n}{2}\right)^{\hat{\alpha}_n} = \left(1+\frac{\sqrt{1+4d_n^2}}{2d_n^2}+\frac{1}{2d_n^2}\right)^{\hat{\alpha}_n} = \left(1+\frac{1}{2\delta\sqrt{n}}+o(\tfrac{1}{\sqrt{n}})\right)^{\hat{\alpha}_n}$$
$$= e^{\frac{\alpha}{2\delta}}+o(1).$$

where we shortened $\hat{\alpha}_n := \alpha\sqrt{n}+o(\sqrt{n})$.

For the second limit, note that $a_n = \sqrt{q_n^2+4q_n} = \frac{1}{\delta\sqrt{n}}(1+o(1))$. Hence,

$$\left(\frac{q_n+2-a_n}{q_n+2+a_n}\right)^{\hat{\alpha}_n} = \left(1-\frac{a_n}{1+o(1)}\right)^{\hat{\alpha}_n} = \left(1-\frac{1}{\delta\sqrt{n}}(1+o(1))\right)^{\hat{\alpha}_n} = e^{-\frac{\alpha}{\delta}}+o(1).$$

Similarly for the last limit, we have that

$$\left(\frac{q_n+2-a_n}{q_n+2+a_n}\right)^{\omega(\sqrt{n})} = \left(1-\frac{1}{\delta\sqrt{n}}(1+o(1))\right)^{\omega(\sqrt{n})} \to 0,$$

which concludes the claims.

□

2.3.4 Asymptotic detection of modular structures

2.3.4.1 A two communities bottleneck graph

Proof of Theorem 2.9. Equation (2.22) Let $\nu^{(G)}$ denote the non-normalized measure on a graph G. Let K_n be the complete graph on n vertices. We can express the partition function of $BG_{n,m}$ in terms of the partition functions and the non-normalized measure of rooting events in the complete graphs K_n and K_m.

Let L_n denote the negative graph Laplacian of K_n. The partition function of K_n is given by

$$Z_{K_n}(q) = q(q+n)^{n-1}. \tag{2.45}$$

Let U be a set of vertices of K_n with $|U| = r$ and write $[qI - L_n]_U$ to denote the submatrix of $qI - L_n$ otained by removing all rows and columns corresponding to

vertices in U. Then non-normalized measure of the event that at least all vertices in U are roots in a random rooted forest of K_n is given by

$$\begin{aligned}\nu^{(K_n)}(U \subseteq R_q) &= q^r \det[qI - L_n]_U = q^r \det[(q+r)I - L_{n-r}] \\ &= q^r Z_{K_{n-r}}(q+r) = q^r (q+r)(q+n)^{n-r-1}.\end{aligned} \tag{2.46}$$

For the partition function of $BG_{n,m}$, Lemma A.4 gives us that

$$\begin{aligned}Z_{BG_{n,m}}(q) &= Z_{K_n}(q)Z_{K_m}(q) + \tfrac{w}{q} Z_{K_n}(q)\nu^{(K_m)}(b' \in R_q) + \tfrac{w}{q} Z_{K_m}(q)\nu^{(K_n)}(b \in R_q) \\ &= q\left(q(q+n)(q+m) + w(q+1)(2q+n+m)\right)(q+n)^{n-2}(q+m)^{m-2}.\end{aligned}$$

Equation (2.24)

We can express $U_q(b,b')$ explicitly by using Theorem 2.6 and eqs. (2.22) and (2.45)

$$U_q(b,b') = \frac{Z_{K_n}(q)Z_{K_m}(q)}{Z_{BG_{n,m}}(q)} \tag{2.47}$$

$$= \frac{q(q+n)(q+m)}{q(q+n)(q+m) + w(q+1)(2q+n+m)}. \tag{2.48}$$

The result of Equation (2.24) follows directly from this expression.

Equation (2.23) We will assume that x and x' both belong to the clique of size n, as the other case can be proven similarly. By Lemma A.4 we have that

$$\begin{aligned}\nu^{(BG_{n,m})}(x \leftrightarrow_q x') &= \nu^{(K_n)}(x \leftrightarrow_q x')Z_{K_m}(q) + \tfrac{w}{q}\nu^{(K_n)}(x \leftrightarrow_q x',\ b \in R_q)Z_{K_m}(q) \\ &\quad + \tfrac{w}{q}\nu^{(K_n)}(x \leftrightarrow_q x')\nu^{(K_m)}(b' \in R_q).\end{aligned}$$

By Equations (2.45) and (2.46) it follows that

$$\begin{aligned}U_q^{(BG_{n,m})}(x,x') &= 1 - \frac{\nu^{(K_n)}(x \leftrightarrow_q x')Z_{K_m}(q) + \frac{w}{q}\nu^{(K_n)}(x \leftrightarrow_q x',\ b \in R_q)Z_{K_m}(q)}{\frac{Z_{K_n}(q)Z_{K_m}(q)}{U_q^{(BG_{n,m})}(b,b')}} \\ &\quad + \frac{\frac{w(q+1)}{q(q+m)}\nu^{(K_n)}(x \leftrightarrow_q x')Z_{K_m}(q)}{\frac{Z_{K_n}(q)Z_{K_m}(q)}{U_q^{(BG_{n,m})}(b,b')}} \\ &= 1 - U_q^{(BG_{n,m})}(b,b')\left(\left(1 + \tfrac{w(q+1)}{q(q+m)}\right)\mathbb{P}^{(K_n)}(x \leftrightarrow_{\Phi_q} x')\tfrac{w(q+1)}{q(q+n)}\mathbb{P}^{(K_n)}(x \leftrightarrow_q x' \mid b \in R_q)\right).\end{aligned} \tag{2.49}$$

Let H denote the graph obtained by removing all outgoing edges of b from K_n, while retaining the ingoing edges. By Lemma A.2 it then holds that $\mathbb{P}^{(K_n)}(x \leftrightarrow_q x' \mid b \in R_q) = \mathbb{P}^{(H)}(x \leftrightarrow_q x')$. Let $\mathbb{P}_x$ denote the law of the random walk on H starting at x and τ_q an independent exponential killing time with rate q. Since the hitting time τ_b has an exponential distribution with rate 1, we can identify the random walk on H killed at rate q with a random walk on K_{n-1} killed at rate $q+1$, by killing the random walk when it hits b. By analyzing Wilson's algorithm on H with the first two random

walks starting at x and x', this gives us that

$$\begin{aligned}
&\mathbb{P}^{(K_n)}(x \leftrightarrow_q x' \mid b \in R_q) = \mathbb{P}^{(H)}(x \leftrightarrow_q x') \\
&= \mathbb{P}^{(K_{n-1})}(x \leftrightarrow_{q+1} x') + \mathbb{P}^{(K_{n-1})}(x \nleftrightarrow_{q+1} x')\ \mathbb{P}_x(\tau_b < \tau_q)\ \mathbb{P}_{x'}(\tau_b < \tau_q) \\
&= \mathbb{P}^{(K_{n-1})}(x \leftrightarrow_{q+1} x') + \tfrac{1}{(q+1)^2}\mathbb{P}^{(K_{n-1})}(x \nleftrightarrow_{q+1} x') \\
&= \tfrac{1}{(q+1)^2} + \tfrac{q(q+2)}{(q+1)^2}\mathbb{P}^{(K_{n-1})}(x \leftrightarrow_{q+1} x').
\end{aligned} \tag{2.50}$$

By [7, Theorem 1] we have that

$$\mathbb{P}^{(K_n)}(x \leftrightarrow_q x') \to \begin{cases} 1 \text{ if } q = o(\sqrt{n}) \\ 0 \text{ if } q = \omega(\sqrt{n}), \end{cases} \tag{2.51}$$

which together with Equation (2.50) gives us that

$$\mathbb{P}^{(K_n)}(x \leftrightarrow_q x' \mid b \in R_q) \to \begin{cases} 1 \text{ if } q = o(\sqrt{n}) \\ 0 \text{ if } q = \omega(\sqrt{n}). \end{cases}$$

Assume that $q = o(\sqrt{n})$. Fix a small enough $\varepsilon > 0$. Then for n large enough it holds that $\mathbb{P}^{(K_n)}(x \leftrightarrow x') > 1 - \varepsilon$ and that $\mathbb{P}^{(K_n)}(x \leftrightarrow x' \mid b \in R_q) > 1 - \varepsilon$. By Equations (2.48) and (2.49), this means that for n large enough

$$U_q^{(BG_{n,m})}(x, x') < 1 - U_q^{(BG_{n,m})}(b, b')\left(\left(1 + \tfrac{w(q+1)}{q(q+m)}\right)(1-\varepsilon) + \tfrac{w(q+1)}{q(q+n)}(1-\varepsilon)\right) = \varepsilon. \tag{2.52}$$

If instead $q = \omega(\sqrt{n})$, then analogously we find for large enough n that

$$U_q^{(BG_{n,m})}(x, x') > 1 - \varepsilon.$$

Equation (2.25) Assume that x, x' and b belong to the clique of size n. By again considering the random walk on H, we find that

$$\mathbb{P}^{(K_n)}(x \leftrightarrow b \mid b \in R_q) = \mathbb{P}_x(\tau_b < \tau_q) = \frac{1}{q+1}.$$

So, since $\frac{1}{q+1} \to 0$ for $q = \omega(\sqrt{n})$, the case $q = \omega(\sqrt{n})$ follows analogous to Equation (2.52).

Now assume that $q = o(\sqrt{n})$. Then we have that $\mathbb{P}^{(K_n)}(x \leftrightarrow_q b) \to 1$, so that

$$\begin{aligned}
U_q^{(BG_{n,m})}(x, b) &= 1 - U_q^{(BG_{n,m})}(b, b')\left(\left(1 + \tfrac{w(q+1)}{q(q+m)}\right)\mathbb{P}^{(K_n)}(x \leftrightarrow_q b) + \tfrac{w(q+1)}{q(q+n)}\ \tfrac{1}{q+1}\right) \\
&\sim 1 - U_q^{(BG_{n,m})}(b, b')\left(\left(1 + \tfrac{w(q+1)}{q(q+m)}\right) + \tfrac{w(q+1)}{q(q+n)}\ \tfrac{1}{q+1}\right) \\
&= \frac{wq(q+m)}{q(q+n)(q+m) + w(q+1)(2q+n+m)}.
\end{aligned}$$

This asymptotic expression for $U_q^{(BG_{n,m})}(x, b)$ gives us that

$$U_q^{(BG_{n,m})}(x, b) \to \begin{cases} 0 & \text{if } \begin{cases} q = o(1) \text{ or } (q = o(\sqrt{|C_i|}),\ w = o(m)) \\ \text{or } (q = o(\sqrt{|C_i|}),\ m = o(n)), \end{cases} \\ \frac{c}{1+c} & \text{if } q = \omega(1),\ q = o(\sqrt{n}),\ w = \omega(m),\ m \sim cn \text{ with } c \in (0, 1] \\ 1 & \text{if } q = \omega(\sqrt{n}) \end{cases}$$

Performing the same computation for $U_q(y, b')$ yields the result of Equation (2.25).

Equation (2.26) By Lemma A.1 and eqs. (2.47) and (2.50) it holds that

$$\begin{aligned}
&U_q^{(BG_{n,m})}(x,y)\\
&= 1 - \frac{\nu^{(BG_{n,m})}(x \leftrightarrow_q y,\ (b,b') \in \Phi_q)}{Z_{BG_{n,m}}(q)} - \frac{\nu^{(BG_{n,m})}(x \leftrightarrow_q y,\ (b',b) \in \Phi_q)}{Z_{BG_{n,m}}(q)}\\
&= 1 - \frac{\frac{w}{q}\nu^{(K_n)}(x \leftrightarrow_q b,\ b \in R_q)\ \nu^{(K_m)}(b' \leftrightarrow_q y)}{Z_{BG_{n,m}}(q)}\\
&\quad - \frac{\frac{w}{q}\nu^{(K_n)}(x \leftrightarrow_q b)\ \nu^{(K_m)}(b' \leftrightarrow_q y,\ b' \in R_q)}{Z_{BG_{n,m}}(q)}\\
&= 1 - U_q^{(BG_{n,m})}(b,b')\left(\frac{w}{q}\mathbb{P}^{(K_n)}(x \leftrightarrow_q b,\ b \in R_q)\ \mathbb{P}^{(K_m)}(b' \leftrightarrow_q y)\right.\\
&\quad \left. + \frac{w}{q}\mathbb{P}^{(K_n)}(x \leftrightarrow_q b)\ \mathbb{P}^{(K_m)}(b' \leftrightarrow_q y,\ b' \in R_q)\right)\\
&= 1 - U_q^{(BG_{n,m})}(b,b')\left(\frac{w}{q(q+n)}\mathbb{P}^{(K_m)}(b' \leftrightarrow_q y) + \frac{w}{q(q+m)}\mathbb{P}^{(K_n)}(x \leftrightarrow_q b)\right)\\
&= 1 - \frac{w(q+m)\mathbb{P}^{(K_m)}(b' \leftrightarrow_q y) + w(q+n)\mathbb{P}^{(K_n)}(x \leftrightarrow_q b)}{q(q+n)(q+m) + w(q+1)(2q+n+m)},
\end{aligned}$$

from which the limits in Equation (2.26) follow. □

2.3.4.2 Star graphs: homogeneous case and with implanted communities

Proof of Proposition 2.10. Let us start by providing an expression for the partition function of a complete k-ary tree with homogeneous weights. Let $w \in (0,\infty)$, $h \in \mathbb{N}$, $k \in \mathbb{N}$, and let L be the negative graph Laplacian of the complete k-ary tree with height h and uniform weight w. Define $(\alpha_n)_{n\in\mathbb{N}_0}$ such that $\alpha_0 = q + w$ and $\alpha_{n+1} = q + (k+1)w - \frac{kw^2}{\alpha_n}$ for $n \in \mathbb{N}$. Then the characteristic polynomial of L is given by

$$\det[qI - L] = \left(\prod_{i=0}^{h-1} \alpha_i^{k^{h-i}}\right)\left(q + kw - \frac{kw^2}{\alpha_{h-1}}\right) \tag{2.53}$$

In fact, observe that in the matrix $[qI - L]$ there is a $k^h \times k^h$ diagonal matrix with entries $q + w$ since the leaves are not connected with each other. Call this right lower diagonal matrix D and call the corresponding left upper matrix A, right upper matrix B and left lower matrix C. By Schur's determinant identity, we get $\det[qI - L] = \det[D]\det[A - BD^{-1}C]$. Here, $\det[D] = (q+w)^{k^h}$ since $D = (q+w)I$. This also gives us $D^{-1} = \frac{1}{q+w}I$. Thus, $BD^{-1}C = \frac{1}{q+w}BC$ is a diagonal matrix with lower entries $\frac{kw^2}{q+w}$, on the places of the parents of the leaves, and upper entries 0, on the places of the nodes that are not parents of the leaves (if there are any). If $h = 1$ we see that $A - BD^{-1}C = q + kw - \frac{kw^2}{q+w}$ and we are done. If $h > 1$ we see that $A - BD^{-1}C$ is again a matrix with a right lower diagonal matrix. This time, the entries of the diagonal matrix are $q + (k+1)w - \frac{kw^2}{q+w}$. By iteration of Schur's determinant identity we get the formula in (2.53).

We'll continue by checking the validity of the expressions in (2.27) and (2.28). By applying (2.53) to the homogeneous star graph SG_n we obtain that its partition function is given by

$$Z_{SG_n}(q) = q(q+w)^{n-2}(q+nw) \tag{2.54}$$

Since $d(c,x) = 1$, by (2.12) we have that

$$U_q(c,x) = \frac{q\ Z_{SG_{n-1}}(q)}{Z_{SG_n}(q)} = \frac{q(q+(n-1)w)}{(q+w)(q+nw)}.$$

Similarly, since $d(x,y) = 2$, by Theorem 2.6 we have that

$$U_q(x,y) = \frac{2q\ Z_{SG_{n-1}}(q)\ -\ q^2\ Z_{SG_{n-2}}(q)}{Z_{SG_n}(q)} = \frac{q(q^2+(n+2)wq+2(n-1)w^2)}{(q+w)^2(q+nw)},$$

which finishes the proof of (2.27) and (2.28). The asymptotics in (2.29) follow immediately. □

Proof of Theorem 2.11. The generator matrix L of the community star graph $CSG_{n,k}$ is given by

$$L = \begin{pmatrix} -k-(n-k-1)w & 1 & 1 & \cdots & w & w \\ 1 & -1 & & & & \\ 1 & & -1 & & & \\ \vdots & & & \ddots & & \\ w & & & & -w & \\ w & & & & & -w \end{pmatrix}$$

where the empty places are to be filled with zeros. The characteristic polynomial of this matrix is

$$\det[qI - L] = q(q+w)^{n-k-2}(q+1)^{k-1}[q^2 + ((n-k)w + k + 1)q + nq] \tag{2.55}$$

which can be found by applying Schur's determinant identity as we did in the proof of Proposition 2.10. Hence, the eigenvalues of L are:

$$\lambda_i = \begin{cases} 0 & \text{if } i = 1 \\ -w & \text{if } i = 2, \ldots, n-k-1 \\ -1 & \text{if } i = n-k, \ldots, n-2 \\ -\frac{1}{2}\mu + \frac{1}{2}\delta & \text{if } i = n-1 \\ -\frac{1}{2}\mu - \frac{1}{2}\delta & \text{if } i = n \end{cases}$$

where

$$\mu = (n-k)w + k + 1$$
$$\delta = \sqrt{((n-1)^2 - 2nk + k^2 + 2n - 1)w^2 + k^2 + 2((n-1)k - k^2 - n)w + 2k + 1}.$$

Denote the sets of vertices that are connected to the center vertex c with a weight 1 and w by V_1 and V_w, respectivley. Combining Theorem 2.6 with Equation (2.55) and writing $c = c(n,k,w,q) = ((n-k)w + k + 1)q + nw$ leads to:

$$U_q(c,x) = \begin{cases} \frac{q(q^2-q-w+c)}{(q+1)(q^2+c)} & x \in V_1 \\ \\ \frac{q(q^2-wq-w+c)}{(q+w)(q^2+c)} & x \in V_w \end{cases}$$

and

$$U_q(x,y) = \begin{cases} \frac{q(q^3+2q^2+(c-2)q+2c-2w)}{(q+1)^2(q^2+c)} & x,y \in V_1 \\ \frac{q(q^3+(w+1)q^2+c(w+q+1)-w(w+1))}{(q+1)(q+w)(q^2+c)} & x \in V_1, y \in V_w \\ \frac{q(q^3+2wq^2+(c-2w-4wk)q+2w(c-w))}{(q+w)^2(q^2+c)} & x,y \in V_w \end{cases}$$

From these explicit formulas, letting q and w be as in the statement, the limits in Theorem 2.11 follow. □

2.3.4.3 Playing with degrees and hierarchical weights on trees

Proof of Proposition 2.12. Note that $\mathbb{P}(e \in \Phi_q) \leq \frac{w(e)}{q+w(e)}$, since by Lemma A.1 it holds that

$$1 \geq \mathbb{P}(e \in \Phi_q) + \mathbb{P}(x \in R_q,\ x \nleftrightarrow_q y) = \mathbb{P}(e \in \Phi_q)(1 + \tfrac{q}{w(e)}).$$

Hence, if $q_k = \omega(w_k(x,y))$, then we have that $U_{q_k}(x,y) \to 1$.

Assume that $q_k = o(w_k(x,y))$. Let $\tilde{\mathbb{P}}_x^{(k)}$ denote the law of the discrete-time random walk $\tilde{X}$ on G_k starting at vertex $x \in V_k$ and let $\tilde{\tau}_q$ be a geometric killing time, as defined in Equation (A.6). Let τ_x denote the first hitting time of x by $\tilde{X}$. Let m denote the number of vertices on the x-side of edge (x,y) in G. We will show by induction on m that

$$\tilde{\mathbb{P}}_x^{(k)}(\tau_y < \tilde{\tau}_q) = 1 - \Theta\left(\tfrac{q_k}{w_k(x,y)}\right).$$

If $m = 1$, then x is a leaf in G. It follows that

$$\tilde{\mathbb{P}}_x^{(k)}(\tau_y < \tilde{\tau}_q) = \frac{w_k(x,y)}{q_k + w_k(x,y)} \sim 1 - \tfrac{q_k}{w_k(x,y)}.$$

Assume that $m \geq 2$. Let N_x denote the set of neighbors of x in G. Since the limit $\lim_{k\to\infty} \frac{w_k(e)}{q_k}$ exists for all edges incident to x, we can partition $N_x \setminus \{y\}$ into two parts: the first part $N_x^{\leq} = \{v \in N_x \setminus \{y\} \colon\ w_k(x,v) = \mathcal{O}(q_k)\}$ consists of all neighbors for which the weight of the edge between x and that neighbor has no larger order than q_k; the second part $N_x^{>} = \{v \in N_x \setminus \{y\} \colon\ w_k(x,v) = \omega(q_k)\}$ consists of the remaining neighbors. Then for each $v \in N_x^{>}$ we have that $q_k = o(w_k(x,v))$. For each such v it follows by the induction hypothesis that $\mathbb{P}_v^{(k)}(\tau_x < \tilde{\tau}_q) = 1 - \Theta\left(\frac{q_k}{w_k(x,v)}\right)$. It follows that

$$\begin{aligned} \tilde{\mathbb{P}}_x^{(k)}(\tau_y < \tilde{\tau}_q) &= \frac{w_k(x,y)}{q_k + w_k(x,y) + \sum_{v \in N_x \setminus \{y\}} w_k(x,v)(1 - \tilde{\mathbb{P}}_v^{(k)}(\tau_x < \tilde{\tau}_q))} \\ &= \frac{w_k(x,y)}{w_k(x,y) + \Theta(q_k) + \sum_{v \in N_x^{\leq}} w_k(x_n,v)(1 - \tilde{\mathbb{P}}_v^{(k)}(\tau_x < \tilde{\tau}_q))} \\ &= \frac{w_k(x,y)}{w_k(x,y) + \Theta(q_k)} = 1 - \Theta\left(\frac{q_k}{w_k(x,y)}\right). \end{aligned}$$

Thus we have that $\tilde{\mathbb{P}}_x^{(k)}(\tau_y < \tilde{\tau}_q) = 1 - o(1)$, from which it follows that $U_{q_k}^{(G_k)}(x,y) \to 0$. □

Lemma 2.18 (Parent hitting asymptotics with small q in hierarchical trees of bounded height). *For each $n \in \mathbb{N}$ let $G_n = (V_n, E_n, w_n)$ be a hierarchical tree of height $H = H_n$, see Figure 2.4a. Denote the weight of an edge at height $i \in [H]$ in G_n by $w_i^{(n)}$ and recall that $w_1^{(n)} \leq \ldots \leq w_H^{(n)}$.*

For each $n \in \mathbb{N}$ let y_n be a vertex in G_n at height $h = h_n$ such that $H_n - h_n$ is constant in n. Let x_n denote the parent of y_n. For each vertex v in G_n let $\ell_n(v)$ denote the number of vertices in G_n that have v in their ancestry. Let $(q_n)_{n\in\mathbb{N}}$ be a sequence of rooting parameters such that $q_n = o\left(\frac{w_h^{(n)}}{\ell_n(y)}\right)$.

For each $n \in \mathbb{N}$ let $\tilde{\mathbb{P}}_x^{(n)}$ denote the law of the discrete-time random walk $\tilde{X}$ on G_n starting at vertex $x \in V_n$ and let $\tilde{\tau}_q$ be a geometric killing time, as defined in Equation (A.6). Let τ_x denote the first hitting time of x by $\tilde{X}$. Then as $n \to \infty$ it holds that

$$\tilde{\mathbb{P}}_y^{(n)}(\tau_x < \tilde{\tau}_q) \sim 1 - \frac{q_n \ell_n(y)}{w_h^{(n)}}.$$

Proof. Write $k = H_n - h_n$, which is independent of n. We proceed by induction on k.

For $k = 0$ we have that all vertices y_n are leaves. We then have that $\ell_n(y) = 1$, so that $q_n = o(w_H^{(n)})$. It follows that

$$\mathbb{P}_y^{(n)}(\tau_x < \tilde{\tau}_q) = \frac{w_H^{(n)}}{w_H^{(n)} + q_n} \sim 1 - \tfrac{q_n}{w_H^{(n)}}.$$

Now assume that $k > 0$. Let $C_y^{(n)} \subseteq V_n$ denote the set of child vertices of y_n in G_n. Note that since $k > 0$, we have for all n that $C_y^{(n)}$ is non-empty. For each $n \in \mathbb{N}$ let z_n be a child of y_n. Note that $\frac{w_h^{(n)}}{\ell_n(y)} \leq \frac{w_{h+1}^{(n)}}{\ell_n(z)}$. This means that $q_n = o\left(\frac{w_{h+1}^{(n)}}{\ell_n(z)}\right)$. Thus by the induction hypothesis we then have that

$$\mathbb{P}_z^{(n)}(\tau_y < \tilde{\tau}_q) \sim 1 - \frac{q_n \ell_n(z)}{w_{h+1}^{(n)}}.$$

Since this holds for all possible choices of sequences of children of y_n, Lemma A.8 stated at the end of this section gives us that

$$\sum_{z \in C_y^{(n)}} \mathbb{P}_z^{(n)}(\tau_y < \tilde{\tau}_q) \sim \sum_{z \in C_y^{(n)}} 1 - \frac{q_n \ell_n(z)}{w_{h+1}^{(n)}}. \tag{2.56}$$

Note that for all n it holds that

$$\begin{aligned}\tilde{\mathbb{P}}_y^{(n)}(\tau_x < \tilde{\tau}_q) = \tilde{\mathbb{P}}_y^{(n)}(X_1 = x) + \tilde{\mathbb{P}}_y^{(n)}(X_1 = y)\tilde{\mathbb{P}}_y^{(n)}(\tau_x < \tilde{\tau}_q) \\ + \sum_{z \in C_y^{(n)}} \tilde{\mathbb{P}}_y^{(n)}(X_1 = z)\tilde{\mathbb{P}}_z^{(n)}(\tau_y < \tilde{\tau}_q)\tilde{\mathbb{P}}_y^{(n)}(\tau_x < \tilde{\tau}_q).\end{aligned}$$

Solving this equation gives us that

$$\tilde{\mathbb{P}}_y^{(n)}(\tau_x < \tilde{\tau}_q) = \frac{\tilde{\mathbb{P}}_y^{(n)}(\tilde{X}_1 = x)}{1 - \tilde{\mathbb{P}}_y^{(n)}(\tilde{X}_1 = y) - \sum_{z \in C_y^{(n)}} \tilde{\mathbb{P}}_y^{(n)}(\tilde{X}_1 = z)\tilde{\mathbb{P}}_z^{(n)}(\tau_y < \tilde{\tau}_q)}$$
$$= \frac{w_h^{(n)}}{q_n + w_h^{(n)} + w_{h+1}^{(n)} \sum_{z \in C_y^{(n)}} \left(1 - \tilde{\mathbb{P}}_z^{(n)}(\tau_y < \tilde{\tau}_q)\right)}. \tag{2.57}$$

Since $q_n = o\left(\frac{w_h^{(n)}}{\ell_n(y)}\right)$, we then have that

$$\tilde{\mathbb{P}}_y^{(n)}(\tau_x < \tilde{\tau}_q) = \frac{w_h^{(n)}}{q_n + w_h^{(n)} + w_{h+1}^{(n)} \sum_{z \in C_y^{(n)}} \left(1 - \tilde{\mathbb{P}}_z^{(n)}(\tau_y < \tilde{\tau}_q)\right)}$$
$$\sim \frac{w_h^{(n)}}{q_n + w_h^{(n)} + w_{h+1}^{(n)} \sum_{z \in C_y^{(n)}} \frac{q_n \ell_n(z)}{w_{h+1}^{(n)}}}$$
$$= \frac{w_h^{(n)}}{q_n + w_h^{(n)} + q_n \sum_{z \in C_y^{(n)}} \ell_n(z)} = \frac{w_h^{(n)}}{w_h^{(n)} + q_n \ell_n(y)} \sim 1 - \frac{q_n \ell_n(y)}{w_h^{(n)}}.$$

□

Proof of Theorem 2.13. If d_n is bounded, then the result follows from Proposition 2.12. Hence, we can assume that $d_n \to \infty$ as $n \to \infty$. Since G_n is a complete d_n-ary tree, the number of vertices with y_n in their ancestry is given by $\ell_n(y) = \sum_{i=0}^{k} d_n^i$. This means that $\ell_n(y) \sim d_n^k$ as as $n \to \infty$. Hence, the case $q_n = o\left(\frac{w_n(e_n)}{d_n^k}\right)$ follows directly from Lemmas 2.16 and 2.18.

Assume that $q_k = \omega\left(\frac{w_k(e_k)}{d_k^m}\right)$. By Theorem 2.3 we can assume without loss of generality that also $q_n = o\left(\frac{w_n(e_n)}{d_n^{k-1}}\right)$.

For ech $n \in \mathbb{N}$ let $\tilde{\mathbb{P}}_x^{(n)}$ denote the law of the discrete-time random walk $\tilde{X}$ on G_n starting at vertex $x \in V_n$ and let $\tilde{\tau}_q$ be a geometric killing time, as defined in Equation (A.6). Let τ_x denote the first hitting time of x by $\tilde{X}$. By Lemma 2.16 it is sufficient to show that both $\tilde{\mathbb{P}}_y^{(n)}(\tau_x < \tilde{\tau}_q) \to 0$ and $\tilde{\mathbb{P}}_x^{(n)}(\tau_y < \tilde{\tau}_q) \to 0$ as $n \to \infty$.

First we consider $\tilde{\mathbb{P}}_y^{(n)}(\tau_x < \tilde{\tau}_q)$. Let z_k be a child vertex of y_k. Then by Lemma 2.18 we have that

$$\tilde{\mathbb{P}}_z^{(n)}(\tau_y < \tilde{\tau}_q) \sim 1 - \frac{q_n \sum_{i=0}^{m-1} d_n^i}{w_n(y_n, z_n)}$$

So, by using that G_n is a complete d-ary tree, we have analogous to Equation (2.57) that

$$\tilde{\mathbb{P}}_y^{(n)}(\tau_x < \tilde{\tau}_q) = \frac{w_n(e_n)}{q_n + w_n(e_n) + d_n \left(1 - \tilde{\mathbb{P}}_z^{(n)}(\tau_y < \tilde{\tau}_q)\right) w_n(y_n, z_n)}$$
$$\sim \frac{w_n(e_n)}{q_n + w_n(e_n) + d_n q_n \sum_{i=0}^{m-1} d_n^i} = \frac{w_n(e_n)}{q_n + w_n(e_n) + \omega(w_n(e_n))} = o(1).$$

It remains to show that $\tilde{\mathbb{P}}_x^{(n)}(\tau_y < \tilde{\tau}_q) \to 0$. Let u denote the parent of x. Then it holds that

$$\begin{aligned}\tilde{\mathbb{P}}_x^{(n)}(\tau_y < \tilde{\tau}_q) = \tilde{\mathbb{P}}_x^{(n)}(\tilde{X}_1 = y) + \tilde{\mathbb{P}}_x^{(n)}(\tilde{X}_1 = x)\tilde{\mathbb{P}}_x^{(n)}(\tau_y < \tilde{\tau}_q) \\ + \tilde{\mathbb{P}}_x^{(n)}(\tilde{X}_1 = u)\tilde{\mathbb{P}}_u^{(n)}(\tau_x < \tilde{\tau}_q)\tilde{\mathbb{P}}_x^{(n)}(\tau_y < \tilde{\tau}_q) \\ + (d_n - 1)\tilde{\mathbb{P}}_x^{(n)}(\tilde{X}_1 = y)\mathbb{P}_y^{(k,q)}(\tau_x < \tilde{\tau}_q)\tilde{\mathbb{P}}_x^{(n)}(\tau_y < \tilde{\tau}_q).\end{aligned}$$

This gives us that $\tilde{\mathbb{P}}_x^{(n)}(\tau_y < \tilde{\tau}_q)$ is equal to

$$\begin{aligned}&\frac{\tilde{\mathbb{P}}_x^{(n)}(\tilde{X}_1 = y)}{1 - \tilde{\mathbb{P}}_x^{(n)}(\tilde{X}_1 = x) - \tilde{\mathbb{P}}_x^{(n)}(\tilde{X}_1 = u)\tilde{\mathbb{P}}_u^{(n)}(\tau_x < \tilde{\tau}_q) - (d_n - 1)\tilde{\mathbb{P}}_x^{(n)}(\tilde{X}_1 = y)\tilde{\mathbb{P}}_y^{(n)}(\tau_x < \tilde{\tau}_q)\tilde{\mathbb{P}}_x^{(n)}(\tau_y < \tilde{\tau}_q)} \\ &= \frac{w_n(e_n)}{q_n + (1 - \tilde{\mathbb{P}}_x^{(n)}(\tau_y < \tilde{\tau}_q))w_n(x,u) + w_n(e_n) + w_n(e_n)(d_n - 1)\left(1 - \tilde{\mathbb{P}}_y^{(n)}(\tau_x < \tilde{\tau}_q)\right)}.\end{aligned}$$

Since we have already shown that $\tilde{\mathbb{P}}_y^{(n)}(\tau_x < \tilde{\tau}_q) \to 0$, it follows that

$$\begin{aligned}\tilde{\mathbb{P}}_x^{(n)}(\tau_y < \tilde{\tau}_q) &\leq \frac{w_n(e_n)}{w_n(e_n) + w_n(e_n)(d_n - 1)\left(1 - \tilde{\mathbb{P}}_y^{(n)}(\tau_x < \tilde{\tau}_q)\right)} \\ &= \frac{1}{1 + (d_n - 1)\,(1 - o(1))} = o(1).\end{aligned}$$

□

APPENDIX A

Appendix: Chapter 2

A.1 Graph reduction/extension lemmas

We introduce here some rather classical contraction tools. Though, we stress that the following definition of contraction is slightly different from what is often encountered in the UST literature, as it is adapted to the setting of weighted directed graphs.

Definition A.1.1 (Directed edge contraction). Let $G = (V, E, w)$ be a weighted directed graph and $e \in E$ a directed edge from vertex x to y, i.e. $e = (x, y)$. The graph $G\vec{/}e$ obtained by performing *the directed edge contraction* in G over edge e is the graph obtained by first removing all outgoing edges of x and then contracting x and y into a single vertex, while retaining all outgoing edges from y and all ingoing edges to both x and y.

If B is a set of edges that constitutes a rooted forest of G, then the operations of performing a directed edge contraction on different edges in B commute. Thus for such a B we can define the graph $G\vec{/}B$ to be the graph obtained by performing directed edge contractions on all edges in B. ■

Besides this notation for directed edge contractions, we will also use the standard notation $G - e$ to denote the graph obtained by removing the directed edge e (without removing the reversed edge), and G/e to denote a regular edge contraction over edge e, i.e. G/e is the graph obtained by identifying the two endpoints of e as a single vertex.

Lemma A.1 (Various expressions for edge probabilities). *Let $G = (V, E, w)$ be a weighted directed graph and $e = (x, y)$ a directed edge from vertex x to y. Let R_q be the set of root vertices of Φ_q. Let L denote the negative graph Laplacian of G and for $q > 0$ write $K_q = q(qI - L)^{-1}$. For each directed edge e write $G\vec{/}e$ to denote the directed e-contraction of G. Then it holds that*

$$\mathbb{P}(e \in \Phi_q) = \frac{w(e)}{q}\mathbb{P}(x \in R_q,\ x \nleftrightarrow_{\Phi_q} y) = \frac{w(e)}{q}(K_q(x,x) - K_q(y,x)) = w(e)\frac{Z_{G\vec{/}e}(q)}{Z_G(q)}.$$

Proof. Let $e = (x, y)$ be an edge from x to y. Let $\mathcal{A} = \{F \in \mathcal{F}_G \colon e \in F\}$ denote the set of rooted forests of G that do contain edge e. Write $\mathcal{H} = \{F \in \mathcal{F}_G \colon x \in R(F),\ x \nleftrightarrow_F y\}$ to denote the set of forests in which x is a root that is not connected to y. Note that there is a one-to-one correspondence $f : \mathcal{A} \to \mathcal{H}$ given by $f(F) = F - e$. Moreover,

it holds that $w(F) = w(e)w(f(F))$ and that $r(F) = r(f(F)) - 1$, where $r(F)$ denotes the number of roots of F. The first identity follows by summation over all forests in $\mathcal{A}$. For the second identity we use the Chebotarev-Shamis matrix-forest theorem [19], which states that $K_q(y, x) = \mathbb{P}(x \in R_q,\ x \leftrightarrow_{\Phi_q} y)$. The third identity follows by considering the bijection $g : \mathcal{H} \to \mathcal{F}_{G\vec{/}e}$ that sends all edges of a forest in $\mathcal{H}$ to their corresponding edges in $G\vec{/}e$. Note that here $G\vec{/}e$ could be a multigraph. This bijection satisfies $w(F) = w(g(F))$ and $r(F) = r(g(F)) + 1$, so that summation over all forests in $\mathcal{H}$ yields the result. □

The following lemma shows the well-known spatial Markov property for the UST, see e.g. [39], tailored to the rooted forest measure Φ_q.

Lemma A.2 (Directed Spatial Markov property). *Let $G = (V, E, w)$ be a weighted directed graph and $A, B \subseteq E$ two disjoint sets of directed edges. Then it holds for all $F \in \mathcal{F}_G$ with $F \cap A = \varnothing$ and $B \subseteq F$ that*

$$\mathbb{P}^{(G)}(\Phi_q = F \mid \Phi_q \cap A = \varnothing,\ B \subseteq \Phi_q) = \mathbb{P}^{((G-A)\vec{/}B)}(\Phi_q = F\vec{/}B). \tag{A.1}$$

For any edge $e \in E$ the partition function of G satisfies the deletion-contraction identity [76]

$$Z_G(q) = Z_{G-e}(q) + w(e)Z_{G\vec{/}e}(q). \tag{A.2}$$

Moreover, if G is a symmetric graph, then it holds that

$$\mathbb{P}^{(G)}(\Phi_q = F \mid \Phi_q \cap A = \varnothing, \pm B \subseteq \Phi_q) = \mathbb{P}^{((G-A)/B)}(\Phi_q = F/B), \tag{A.3}$$

where G/B denotes the regular edge contraction of all edges in B.

Proof. It is sufficient to show that the statement holds when $|A \cup B| = 1$, since the general statement then follows by induction. First assume that $B = \varnothing$ and $A = \{e\}$ for some edge $e \in E$. Let $\mathcal{A} = \{F \in \mathcal{F}_G \colon e \notin F\}$ denote the set of rooted forests of G that do not contain edge e. Write $r(F)$ to denotes the number of roots of the rooted forest F. There is a natural one-to-one correspondence $f : \mathcal{A} \to \mathcal{F}_{G-e}$ given by $f(F) = F$. Hence, we have for all $F \in \mathcal{A}$ that

$$\begin{aligned}
\mathbb{P}^{(G)}(\Phi_q = F \mid e \notin \Phi_q) &= \frac{\mathbb{P}^{(G)}(\Phi_q = F)}{\mathbb{P}^{(G)}(e \notin \Phi_q)} = \frac{q^{r(F)}w(F)}{\sum_{H\in\mathcal{A}} q^{r(H)}w(H)} \\
&= \frac{q^{r(F)}w(F)}{\sum_{H\in\mathcal{F}_{G-e}} q^{r(f^{-1}(H))}w(f^{-1}(H))} \\
&= \frac{q^{r(F)}w(F)}{\sum_{H\in\mathcal{F}_{G-e}} q^{r(H)}w(H)} = \frac{q^{r(F)}w(F)}{Z_{G-e}(q)} = \mathbb{P}^{(G-e)}(\Phi_q = F).
\end{aligned}$$

Assume instead that $A = \varnothing$ and $B = \{e\}$ for some edge $e \in E$. Then by lemma A.1 we have for all $F \in \mathcal{F}_G$ with $e \in F$ that

$$\mathbb{P}^{(G)}(\Phi_q = F \mid e \in \Phi_q) = \frac{\mathbb{P}^{(G)}(\Phi_q = F)}{\mathbb{P}^{(G)}(e \in \Phi_q)} = \frac{q^{r(F)}w(F)}{w(e)Z_{G\vec{/}e}(q)} = \mathbb{P}^{(G\vec{/}e)}(\Phi_q = F/e).$$

The proof of eq. (A.2) is analogous to that of eq. (A.1), while eq. (A.3) follows directly from the spatial Markov property for the UST. □

Lemmas A.3 and A.4 both represent the same simple combinatorial manipulation, but in two slightly different settings. The same manipulation can be extended beyond the simple setups of these lemmas, but for notational simplicity we stick to these versions, which are tailored to sparse geometries.

These lemmas are phrased in terms of the *non-normalized* rooted forest measure defined as

$$\nu^{(G)}(\Phi_q \in \cdot) = Z_G(q)\mathbb{P}^{(G)}(\Phi_q \in \cdot). \tag{A.4}$$

This measure has the benefit that the measure of a rooted forest depends on the geometry of the underlying graph only through the total number of vertices. That is, for any rooted forest $F \in \mathcal{F}_H$ of a subgraph H of G it holds that $q^m \nu^{(H)}(\Phi_q = F) = \nu^{(G)}(\Phi_q = F)$, where m is the difference between the number of vertices in G and H. This simplifies the notation required for various combinatorial manipulations.

Lemma A.3 (Graph extension lemma (single vertex version)). *Let $G = (V, E, w)$ be a weighted directed graph and $x \in V$ a vertex. Let R_q be the set of root vertices of Φ_q. Let $H = G[V \setminus \{x\}]$ denote the induced subgraph of G obtained by removing vertex x. Let $\{\mathcal{H}(F)\colon\ F \in \mathcal{F}_H\}$ be the partition of $\mathcal{F}_G$ given by $\mathcal{H}(F) = \{F' \in \mathcal{F}_G\colon\ F'[V \setminus \{x\}] = F\}$, i.e. $\mathcal{H}(F)$ denotes the set of spanning rooted forests of G for which the induced subgraph obtained by removing x equals F. For each vertex $y \in V \setminus \{x\}$ let $r_y(F)$ denote the unique root in F that is connected to y. Then it holds for all $F \in \mathcal{F}_H$ that*

$$\nu^{(G)}(\Phi_q \in \mathcal{H}(F),\ x \in R_q) = q\ \nu^{(H)}(\Phi_q = F) \prod_{r \in R(F)} (1 + \tfrac{w(r,x)}{q})$$

and that

$$\nu^{(G)}(\Phi_q \in \mathcal{H}(F),\ x \notin R_q) = \nu^{(H)}(\Phi_q = F) \sum_{y \in V \setminus \{x\}} w(x,y) \prod_{r \in R(F) \setminus \{r_y(F)\}} (1 + \tfrac{w(r,x)}{q}).$$

Here we take $w(e) = 0$ when $e \notin E$.

Proof of lemma A.3. We will first prove the first equality. Let $F_H \in \mathcal{F}_H$ be given. Each forest in $F \in \mathcal{H}(F_H)$ with $x \in R(F)$ can be obtained from F_H by adding any number of edges from roots of F_H to x. So, for each root we can choose either to add this edge or not to add this edge. For each edge we do add, there will be one less component, since the root from which that edge originated will cease to be a root in the new forest. This contributes a factor $\frac{1}{q}$. We then also have an additional edge, which contributes a factor equal to the weight of that edge. This gives us the product over the roots r, where the 1 term is chosen if no edge is added from r to x and the $\frac{w(r,x)}{q}$ term is chosen if we do add such an edge. If we don't add any such edges, then the obtained forest will have one more root than F_H, so this gives us the additional factor q.

The second equality is proven similarly. Each forest in $F \in \mathcal{H}(F_H)$ with $x \notin R(F)$ can be obtained from F_H by first adding a single edge from x to any other vertex y. We then add any number of edges from roots of F_H to x, but we cannot add an edge from r_y to x as this would create a cycle. □

Definition A.1.2. Let $G = (V, E)$ be a directed graph. Let $A \subseteq V$ be a set of vertices and denote by $G[A]$ the induced subgraph of G on the vertices in A. A set $\mathcal{H} \subseteq \mathcal{F}$ of rooted forests of G is said to be *determined* by A if there exists an $\mathcal{A} \subseteq \mathcal{F}_{G[A]}$ such that $\mathcal{H} = \{F \in \mathcal{F}_G \colon\ F[A] \in \mathcal{A}\}$. ■

Lemma A.4 (Graph extension lemma (single edge version)). *Let $G = (V, E, w)$ be a weighted directed graph. Let $\{A, B\}$ be a partition of V and assume that there exists exists exactly one vertex $a \in A$ that is adjacent to any vertices in B and exactly one vertex $b \in B$ adjacent to any vertices in A. Write $G[A]$ and $G[B]$ to denote the induced subgraphs on A and B. Let $\mathcal{A}, \mathcal{B} \subseteq \mathcal{F}$ be sets of rooted forests of G that are determined by A and B, respectively, and let $\mathcal{A}' \subseteq \mathcal{F}_{G[A]}$ and $\mathcal{B}' \subseteq \mathcal{F}_{G[B]}$ be such that $\mathcal{A} = \{F \in \mathcal{F}_G \colon\ F[A] \in \mathcal{A}'\}$ and $\mathcal{B} = \{F \in \mathcal{F}_G \colon\ F[B] \in \mathcal{B}'\}$. Denote by R_q the set of root vertices of Φ_q. Then it holds that*

$$\begin{aligned}\nu^{(G)}(\Phi_q \in \mathcal{A} \cap \mathcal{B}) = {} & \nu^{(G[A])}(\Phi_q \in \mathcal{A}')\ \nu^{(G[B])}(\Phi_q \in \mathcal{B}') \\ & + \tfrac{w(a,b)}{q} \nu^{(G[B])}(\Phi_q \in \mathcal{B}')\ \nu^{(G[A])}(\Phi_q \in \mathcal{A}',\ a \in R_q) \\ & + \tfrac{w(b,a)}{q} \nu^{(G[A])}(\Phi_q \in \mathcal{A}')\ \nu^{(G[B])}(\Phi_q \in \mathcal{B}',\ b \in R_q).\end{aligned}$$

Proof. Let $F_A \in \mathcal{A}'$ and $F_B \in \mathcal{B}'$ be given.

If both a is a root in F_A and b is a root in F_B, then there are exactly three forests $F_1, F_2, F_3 \in \mathcal{F}_G$ for which the induced subgraphs on A and B correspond to F_A and F_B, respectively.

(a) The first of these forests consists of the disjoint graph union of F_A and F_B and has non-normalized measure

$$\nu^{(G)}(\Phi_q = F_1) = \nu^{(G[A])}(\Phi_q = F_A)\ \nu^{(G[B])}(\Phi_q = F_B).$$

(b) The second has an additional edge from a to b and has non-normalized measure

$$\nu^{(G)}(\Phi_q = F_2) = \tfrac{w(a,b)}{q} \nu^{(G[A])}(\Phi_q = F_A)\ \nu^{(G[B])}(\Phi_q = F_B),$$

since it contains one less root than the sum of the roots in F_A and F_B and one additional edge with weight $w(a, b)$.

(c) The third forest has an additional edge from b to a and it similarly has non-normalized measure

$$\nu^{(G)}(\Phi_q = F_3) = \tfrac{w(b,a)}{q} \nu^{(G[A])}(\Phi_q = F_A)\ \nu^{(G[B])}(\Phi_q = F_B).$$

Note that each of these three forests is contained in $\mathcal{A} \cap \mathcal{B}$.

If exactly one of the vertices a and b is a root in F_A and F_B, then only two of the above mentioned forests are rooted forest of G, since adding an outgoing edge to a non-root vertex does not yield a rooted forest.

If both a and b are not roots, then only the first forest without an additional edge is a rooted forest of G.

Since each forest in $\mathcal{A} \cap \mathcal{B}$ can be obtained in such a manner, summing over all rooted forests in $\mathcal{A}'$ and $\mathcal{B}'$ yields

$$\begin{aligned}\nu_q^{(G)}(\mathcal{A} \cap \mathcal{B}) = \sum_{F_A \in \mathcal{A}'} \sum_{F_B \in \mathcal{B}'} & \nu^{(G[A])}(\Phi_q = F_A)\nu^{(G[B])}(\Phi_q = F_B) \\ & + \tfrac{w(a,b)}{q}\nu^{(G[A])}(\Phi_q = F_A)\nu^{(G[B])}(\Phi_q = F_B)\mathbf{1}_{\{b \in R(F_A)\}} \\ & + \tfrac{w(b,a)}{q}\nu^{(G[A])}(\Phi_q = F_A)\nu^{(G[B])}(\Phi_q = F_B)\mathbf{1}_{\{b' \in R(F_B)\}} \\ = \nu^{(G[A])}(\Phi_q \in \mathcal{A}')\ & \nu^{(G[B])}(\Phi_q \in \mathcal{B}') \\ + \tfrac{w(a,b)}{q}\nu^{(G[A])}(\Phi_q \in \mathcal{A}',\ a \in R_q)\ & \nu^{(G[B])}(\Phi_q \in \mathcal{B}') \\ + \tfrac{w(b,a)}{q}\nu^{(G[A])}(\Phi_q \in \mathcal{A}')\ & \nu^{(G[B])}(\Phi_q \in \mathcal{B}',\ b \in R_q).\end{aligned}$$

□

A.2 Technical lemmas

The next three lemmas are simple statements used in the proof of theorem 2.5 in section 2.3.2.1.

Lemma A.5 (Bound on derivative of hitting probabilities). *Let $G = (V, E, w)$ be a weighted directed graph and $x, y \in V$ two vertices. Let $\mathbb{P}_x$ denote the law of the random walk X on G starting at x. For each $v \in V$ let τ_v denote the hitting time of v by X and let τ_q be an independent exponential killing time with rate q. Then it holds for the derivative of the function $q \mapsto \mathbb{P}_x(\tau_y < \tau_q)$ that*

$$\tfrac{1}{q}\mathbb{P}_x(\tau_y < \tau_q) - \tfrac{1}{q} \leq \frac{d}{dq}\mathbb{P}_x(\tau_y < \tau_q) \leq 0. \tag{A.5}$$

In the subsequent proofs it will be convenient to work with the discrete-time skeleton of the random walk X, that is, the discrete-time random walk $\tilde{X}$ on G with transition matrix

$$P = I + \tfrac{1}{\alpha}L, \tag{A.6}$$

with α the maximal diagonal entry of the graph Laplacian $-L$. The path measure of $\tilde{X}$ starting at x is denoted by $\tilde{\mathbb{P}}_x$. For $\tilde{\tau}_q$ an independent ($\mathbb{N}$-valued) geometric killing time with success probability $\frac{q}{q+\alpha}$, it then holds that $\mathbb{P}_x(\tau_y < \tau_q) = \tilde{\mathbb{P}}_x(\tau_y < \tilde{\tau}_q)$. Since the law of the loop-erased trajectory of $\tilde{X}$ corresponds to that of X, we can also use this discrete-time random walk to analyze eq. (2.8).

Proof of lemma A.5. The upper bound on the derivative in (A.5) is immediate, we therefore show the lower bound.

Let $\tilde{\mathbb{P}}_x$ denote the law of the discrete-time random walk $\tilde{X}$, as defined in eq. (A.6).

Then it holds that

$$\begin{aligned}\mathbb{P}_x(\tau_y < \tau_q) = \tilde{\mathbb{P}}_x(\tau_y < \tilde{\tau}_q) &= \sum_{k=1}^{\infty} \tilde{\mathbb{P}}_x(\tau_y < k)\mathbb{P}(\tilde{\tau}_q = k) \\ &= \sum_{k=1}^{\infty} \tilde{\mathbb{P}}_x(\tau_y < k)\tfrac{q}{q+\alpha}\left(1 - \tfrac{q}{q+\alpha}\right)^{k-1}.\end{aligned}$$

Since $\tilde{\mathbb{P}}_x(\tau_y < k)$ does not depend on q, it follows that

$$\begin{aligned}\frac{d}{dq}\mathbb{P}_x(\tau_y < \tau_q) &= \sum_{k=1}^{\infty} \tilde{\mathbb{P}}_x(\tau_y < k)\frac{(1-k)q+\alpha}{(q+\alpha)^2}\left(\frac{\alpha}{q+\alpha}\right)^{k-1} \\ &= \tfrac{1}{q}\mathbb{P}_x(\tau_y < \tau_q) - \sum_{k=1}^{\infty} \tilde{\mathbb{P}}_x(\tau_y < k)\frac{kq}{(q+\alpha)^2}\left(\frac{\alpha}{q+\alpha}\right)^{k-1} \\ &\geq \tfrac{1}{q}\mathbb{P}_x(\tau_y < \tau_q) - \sum_{k=1}^{\infty} \frac{kq}{(q+\alpha)^2}\left(\frac{\alpha}{q+\alpha}\right)^{k-1} \\ &= \tfrac{1}{q}\mathbb{P}_x(\tau_y < \tau_q) - \tfrac{1}{q+\alpha}\mathbb{E}(\tilde{\tau}_q) = \tfrac{1}{q}\mathbb{P}_x(\tau_y < \tau_q) - \tfrac{1}{q}.\end{aligned}$$

□

Lemma A.6 (Monotonicity of rooting probabilities). *Let $G = (V, E, w)$ be a weighted directed graph and $x \in V$ a vertex. Let $\mathbb{P}$ denote the law of a random spanning rooted forest Φ_q of G with rooting parameter $q > 0$. Let R_q denote the set of roots of Φ_q. Then it holds that*

$$0 \leq \frac{d}{dq}\mathbb{P}(x \in R_q) \leq \tfrac{1}{q}\mathbb{P}(x \in R_q). \tag{A.7}$$

Proof of lemma A.6 via lemma A.5. By (2.6) we have that x is a root in Φ_q if

$$\mathbb{P}(x \in R_q) = K_q(x,x) = q(qI - L)^{-1}(x,x) = \mathbb{P}_x(X_{\tau_q} = x).$$

Let N_x denote the set of out-neighbours of x in G. Let $\sigma = \inf\{t > 0 \colon X_t \neq X_0\}$ be the first jump time of X. Then by the Markov property of X we have that

$$\mathbb{P}_x(X_{\tau_q} = x) = \mathbb{P}_x(\sigma > \tau_q) + \sum_{v \in N_x} \mathbb{P}_x(X_\sigma = v)\mathbb{P}_v(\tau_x < \tau_q)\mathbb{P}_x(X_{\tau_q} = x).$$

Solving this equation gives us that

$$\begin{aligned}\mathbb{P}_x(X_{\tau_q} = x) &= \frac{\mathbb{P}_x(\sigma > \tau_q)}{1 - \sum_{v \in N_x} \mathbb{P}_x(X_\sigma = v)\mathbb{P}_v(\tau_x < \tau_q)} \\ &= \frac{q}{q + \sum_{v \in N_x} w(x,v)(1 - \mathbb{P}_v(\tau_x < \tau_q))}.\end{aligned}$$

It follows by lemma A.5 that

$$\begin{aligned}\frac{d}{dq}\mathbb{P}(x \in R_q) &= \frac{d}{dq}\mathbb{P}_x(X_{\tau_q} = x) \\ &= \frac{\sum_{v \in N_x} w(x,v)\left(1 - \mathbb{P}_v(\tau_x < \tau_q) + q\frac{d}{dq}\mathbb{P}_v(\tau_x < \tau_q)\right)}{\left(q + \sum_{v \in N_x} w(x,v)(1 - \mathbb{P}_v(\tau_x < \tau_q))\right)^2} \geq 0,\end{aligned}$$

which proves the lower bound. For the upper bound it holds that

$$\frac{\frac{d}{dq}\mathbb{P}(x \in R_q)}{\mathbb{P}(x \in R_q)} = \frac{\sum_{v \in N_x} w(x,v)\left(1 - \mathbb{P}_v(\tau_x < \tau_q) + q\frac{d}{dq}\mathbb{P}_v(\tau_x < \tau_q))\right)}{q\left(q + \sum_{v \in N_x} w(x,v)(1 - \mathbb{P}_v(\tau_x < \tau_q))\right)} \leq \frac{\sum_{v \in N_x} w(x,v)\left(1 - \mathbb{P}_v(\tau_x < \tau_q)\right)}{q\left(q + \sum_{v \in N_x} w(x,v)(1 - \mathbb{P}_v(\tau_x < \tau_q))\right)} \leq \frac{1}{q}.$$

□

Lemma A.7 (Bound on conditional rooting derivative in trees). *Let $G = (V, E, w)$ be a weighted directed tree and $x, y \in V$ two vertices. Then it holds that*

$$\tfrac{d}{dq}\mathbb{P}(x \in R_q \mid x \leftrightarrow y) \leq \tfrac{1}{q}\mathbb{P}(x \in R_q \mid x \leftrightarrow y). \tag{A.8}$$

Proof of lemma A.7. Let d denote the distance between x and y. We will argue inductively on d. For $d = 0$ the statement follows from lemma A.6.

Now assume that $d \geq 1$. Let z denote the vertex adjacent to x with distance $d - 1$ to y. Note that we possibly have that $z = y$. Since G is a tree, removing the edges between x and z splits the graph into two components T_x and T_z, where T_x and T_z denote the component containing vertex x and z, respectively. It then holds by lemma A.4 that

$$\begin{aligned}
&\mathbb{P}(x \in R_q \mid x \leftrightarrow y) \\
&= \frac{w(z,x)\nu^{(T_x)}(x \in R_q)\nu^{(T_z)}(z \in R_q, z \leftrightarrow y)}{w(z,x)Z_{T_x}(q)\nu^{(T_z)}(z \in R_q, z \leftrightarrow y) + w(x,z)\nu^{(T_x)}(x \in R_q)\nu^{(T_z)}(z \leftrightarrow y)} \\
&= \frac{w(z,x)\mathbb{P}^{(T_z)}(z \in R_q \mid z \leftrightarrow y)\mathbb{P}^{(T_x)}(x \in R_q)}{w(z,x)\mathbb{P}^{(T_z)}(z \in R_q \mid z \leftrightarrow y) + w(x,z)\mathbb{P}^{(T_x)}(x \in R_q)}.
\end{aligned}$$

It follows by the induction hypothesis and lemma A.6 that

$$\begin{aligned}
&\frac{d}{dq}\mathbb{P}(x \in R_q \mid x \leftrightarrow y)/\mathbb{P}(x \in R_q \mid x \leftrightarrow y) \\
&= \frac{w(z,x)\mathbb{P}^{(T_x)}(x \in R_q)^2 \frac{d}{dq}\mathbb{P}^{(T_z)}(z \in R_q \mid z \leftrightarrow y) + w(x,z)\mathbb{P}^{(T_z)}(z \in R_q \mid z \leftrightarrow y)^2 \frac{d}{dq}\mathbb{P}^{(T_x)}(x \in R_q)}{w(z,x)\mathbb{P}^{(T_z)}(z \in R_q \mid z \leftrightarrow y)\mathbb{P}^{(T_x)}(x \in R_q)^2 + w(x,z)\mathbb{P}^{(T_z)}(z \in R_q \mid z \leftrightarrow y)^2\mathbb{P}^{(T_x)}(x \in R_q)} \\
&\leq \frac{1}{q}.
\end{aligned}$$

□

The simple lemma below has been used to show eq. (2.56).

Lemma A.8. *For each $n \in \mathbb{N}$ let $\ell_n \in \mathbb{N}$ be given and let $(\alpha_i^{(n)})_{i \in [\ell_n]}$ and $(\beta_i^{(n)})_{i \in [\ell_n]}$ be real valued sequences of length ℓ_n. Let $\mathcal{F} = \{f \in \mathbb{N}^{\mathbb{N}} \colon\ f(n) \in [\ell_n] \text{ for all } n \in \mathbb{N}\}$ denote the set of choice functions on the collection $\{[\ell_1], [\ell_2], \ldots\}$. Assume that for each $f \in \mathcal{F}$ it holds that $\alpha_{f(n)}^{(n)} \sim \beta_{f(n)}^{(n)}$ as $n \to \infty$. Then as $n \to \infty$ it holds that*

$$\sum_{i=1}^{\ell_n} \alpha_i^{(n)} \sim \sum_{i=1}^{\ell_n} \beta_i^{(n)}.$$

Proof. For all $\varepsilon > 0$ and each $f \in \mathcal{F}$, there exists an $N(\varepsilon, f) \in \mathbb{N}$ such that for all $n \geq N(\varepsilon, f)$ it holds that

$$\left| \frac{\alpha_{f(n)}^{(n)}}{\beta_{f(n)}^{(n)}} - 1 \right| < \varepsilon.$$

Define the function $f^* \in \mathcal{F}$ by

$$f^*(n) = \text{argmax}_{i \in [\ell_n]} \left| \alpha_i^{(n)} - \beta_i^{(n)} \right|.$$

Then for all $\varepsilon > 0$ and all $n \geq N(\varepsilon, f^*)$ it holds that

$$\begin{aligned} \left| \frac{\sum_{i=1}^{\ell_n} \alpha_i^{(n)}}{\sum_{i=1}^{\ell_n} \beta_i^{(n)}} - 1 \right| &= \frac{1}{\sum_{i=1}^{\ell_n} \beta_i^{(n)}} \left| \sum_{i=1}^{\ell_n} \alpha_i^{(n)} - \sum_{i=1}^{\ell_n} \beta_i^{(n)} \right| \leq \frac{\sum_{i=1}^{\ell_n} \left| \alpha_i^{(n)} - \beta_i^{(n)} \right|}{\sum_{i=1}^{\ell_n} \beta_i^{(n)}} \\ &= \frac{\sum_{i=1}^{\ell_n} \beta_i^{(n)} \left| \frac{\alpha_i^{(n)}}{\beta_i^{(n)}} - 1 \right|}{\sum_{i=1}^{\ell_n} \beta_i^{(n)}} \leq \left| \frac{\alpha_{f^*(n)}^{(n)}}{\beta_{f^*(n)}^{(n)}} - 1 \right| \frac{\sum_{i=1}^{\ell_n} \beta_i^{(n)}}{\sum_{i=1}^{\ell_n} \beta_i^{(n)}} = \left| \frac{\alpha_{f^*(n)}^{(n)}}{\beta_{f^*(n)}^{(n)}} - 1 \right| < \varepsilon. \end{aligned}$$

□

CHAPTER 3

Wilson's occupation field along coupled Kirchhoff forests

This chapter is based on joint work in progress with L. Avena and A. Gaudillière.

3.1 Introduction

Wilson's algorithm is a celebrated procedure, that uses loop-erased random walks to efficiently sample from the well known Uniform Spanning Tree measure, and from the Kirchhoff forest measure. The latter is a distribution on the spanning rooted forests of a given weighted directed graph, and can be seen as a parametric generalization of the Uniform Spanning Tree measure. The number of components (i.e. trees) in the resulting Kirchhoff forest can be tuned by adjusting the intensity parameter $q > 0$ of the Kirchhoff forest measure. For large values of q a Kirchhoff forest will consist of many small trees, while for small q the measure will concentrate on rooted forest with few components. In particular, the Uniform Spanning Tree is recovered in the limit $q \to 0$.

In [6] it is shown that Wilson's algorithm can be used to couple together a continuum of Kirchhoff forests for all possible intensities $q \in (0, \infty)$. By reparametrizing $t := 1/q$ this coupling constructs a Markov process indexed by time t, such that for each time t the time marginal is a Kirchhoff forest of intensity $1/t$. A partial trajectory of this process, with $t \in [0, t_{\max}]$, can be sampled with approximately the same efficiency as sampling a single Kirchhoff forest of intensity $1/t_{\max}$. Recently, this forest coupling has been applied for estimating the spectrum of a graph's Laplacian matrix [9].

This coupling of realizations of Wilson's algorithm also couples together the associated occupation fields, which are obtained from the loops that are removed during Wilson's procedure. Thus is constructed a stochastic process of occupation fields, which will be called *(Wilson's) occupation field process*, defined in section 3.2.1 below, and which is the object of interest in the present chapter.

The occupation field of Wilson's algorithm is of independent interest, as it shows remarkable connections to Poissonian loop-ensembles (or random walk loop soups) and to the discrete Gaussian free field. These connections were first explored by Le Jan [55].

Outline

This chapter is organized as follows. In section 3.1.1 we will introduce the setting of this chapter. A detailed description of Wilson's algorithm will be provided in section 3.1.2. This description is used in section 3.2 to define the Kirchhoff forest coupling and the associated occupation field process.

In section 3.3 we state the results of this chapter. The main result, theorem 3.6, will give a complete description of the law of the occupation field process. A direct consequence of this theorem, corollary 3.6.1, shows how the connections between the occupation field of Wilson's algorithm and random walk loop soup can be extended to this dynamic setting. The proofs of these results will be deferred to section 3.4.

3.1.1 Setting

Given are a finite set $\mathcal{X}$ of size $n := |\mathcal{X}|$, and an arbitrary *Laplacian matrix* $L = (L(x,y))_{x,y\in\mathcal{X}} \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$, i.e. diagonal entries of L are non-negative, its off-diagonal entries are non-positive, and all of its row sums equal zero. Together the set $\mathcal{X}$ and the Laplacian matrix L form the fixed inputs of our model.

On the set $\mathcal{X}$ we define a discrete-time Markov chain $X = (X_k)_{k\in\mathbb{N}_0}$ with transition matrix $I - \frac{1}{\delta}L$, where I is the identity matrix, and δ is an auxiliary parameter that needs to satisfy

$$\delta \geq \max_{x\in\mathcal{X}} L(x,x) \tag{3.1}$$

for the transition matrix to be well-defined. This parameter δ determines the 'lazyness' of the discrete-time Markov chain, so for larger values of δ the chain is more likely to stay in the same state at each time-step.

The transition matrix $I - \frac{1}{\delta}L$ defines a weighted directed graph $\mathcal{G} = (\mathcal{X}, \mathcal{E}, w)$ with vertices $\mathcal{X}$. The weight function $w : \mathcal{X}\times\mathcal{X} \to [0,\infty)$ is defined by

$$w(x,y) := (\delta I - L)(x,y), \tag{3.2}$$

and the directed edge set is given by

$$\mathcal{E} := \{(x,y) \in \mathcal{X}\times\mathcal{X} \colon\ w(x,y) > 0\}. \tag{3.3}$$

Note that the graph $\mathcal{G}$ has self-loops at vertices x with $\sum_{y\in\mathcal{X}\setminus\{x\}} w(x,y) < \delta$. We further introduce the notation $w(E) := \prod_{(x,y)\in E} w(x,y)$ for the weight of a set of edges $E \subseteq \mathcal{E}$. The Markov chain X will be called the *random walk* on $\mathcal{G}$.

Of interest are the *(spanning) rooted forests* of the graph $\mathcal{G}$, which are subsets $F \subseteq \mathcal{E}$ of directed edges such that:

(*i*) each vertex has at most one outgoing edge in F;

(*ii*) F does not contain any undirected cycles.

The *roots* of a rooted forests F are those vertices that do not have an outgoing edge in F. We denote by $\rho(F)$ the set of roots of F, and by $r(F) := |\rho(F)|$ the number of roots. The set of all rooted forests of the graph is denoted by $\mathcal{F}$.

A *Kirchhoff forest* of intensity $q > 0$ is an $\mathcal{F}$-valued random variable Φ_q with distribution

$$\mathbb{P}(\Phi_q = F) := \tfrac{1}{Z(q)} q^{r(F)} w(F), \tag{3.4}$$

where $Z(q) := \sum_{F\in\mathcal{F}} q^{r(F)} w(F)$ denotes the normalizing partition function.

3.1.2 Wilson's algorithm

Wilson's algorithm is a method for sampling Kirchhoff forests, that is not only of interest from an applied point of view, but is also a useful tool for the theoretical analysis of Kirchhoff forests.

A brief description of Wilson's procedure can be given as follows. Define a killed random walk on the graph. Pick an arbitrary vertex and run a killed random walk

starting from that vertex. Each time the random walk makes a cycle, the edges in the cycle are removed from the trajectory of the random walk, to obtain the loop-erased trajectory.
Then repeatedly pick a new vertex that has not been picked before and run a random walk until either is killed or it hits a vertex in the loop-erased trajectory of any of the previous random walks, and continue until all vertices have been picked. The union of all directed edges in the loop-erased trajectories of the random walks now form a Kirchhoff forest. The intensity of the Kirchhoff forest can be tuned by adjusting the killing time of the random walk.

To define the coupled forest process and the associated occupation field process we require a more detailed description of Wilson's procedure. In fact, we will give two distinct, albeit equivalent, such descriptions. The first is given in section 3.1.2.2 and utilizes loop-erased random walks, while the second, given in section 3.1.2.4, uses a 'cycle popping' procedure based on the Diaconis-Fulton stack representation of a random walk.

We note that Wilson's method depends only indirectly on the law of the used random walks, as only the law of the loop-erased trajectories are relevant. As there are multiple killed random walks of which the distributions of their loop-trajectories coincide, there is some freedom in the choice of random walk. The descriptions of Wilson's algorithm given below, will use the random walk X defined in section 3.1.1 above. This specific choice of random walk will be elucidated in section 3.2.1.1.

3.1.2.1 Notation for walks on graphs

Let $\mathcal{P}$ denote the set of all finite length *walks* in $\mathcal{G}$, i.e.

$$\mathcal{P} := \bigcup_{l\in\mathbb{N}_0} \{(x_0, x_1, \dots, x_l) \in \mathcal{X}^{l+1} \colon\ (x_{i-1}, x_i) \in \mathcal{E} \text{ for all } i \in [l]\}, \tag{3.5}$$

where we use the notation $[l] := \{1, 2, \dots, l\}$ for the set containing the first l positive integers. Elements of $\mathcal{P}$ will commonly be denoted by $\gamma = (x_0, x_1, \dots, x_l)$. For $\gamma \in \mathcal{P}$ its *length* is the unique non-negative integer $l \in \mathbb{N}_0$ for which $\gamma \in \mathcal{X}^{l+1}$. So in particular single vertices are walks of zero length. We introduce the notations

$$s(\gamma) := \{x_0, x_1, \dots, x_l\}, \text{ and } e(\gamma) := \{(x_{i-1}, x_i) \in \mathcal{E} \colon\ i \in [l]\} \tag{3.6}$$

for the support of γ, and the set of edges traversed by γ, respectively. If γ has length zero, then its set of traversed edges is empty. A walk $(x_0, x_1, \dots, x_l) \in \mathcal{P}$ with length $l \geq 1$ for which $x_0 = x_l$ is called a *closed walk* or *cyclic walk*. The set of closed walks is denoted by $\mathcal{P}^{\mathrm{cl}}$. A *self-avoiding walk* is a walk for which $x_i \neq x_j$ for all distinct $i, j \in [l]_0 := \{0, 1, \dots, l\}$.

Define an an equivalence relation $\simeq$ on the set of closed walks $\mathcal{P}^{\mathrm{cl}}$ as follows. For $\gamma_1, \gamma_2 \in \mathcal{P}^{\mathrm{cl}}$ with $\gamma_1 = (x_0, \dots, x_l)$ and $\gamma_2 = (y_0, \dots, y_l)$ we call γ_1 and γ_2 equivalent if there exists a cyclic permutation $\sigma : [l-1]_0 \to [l-1]_0$ such that $x_{\sigma(i)} = y_i$ for all $i \in [l-1]_0$.

We let γ° denote an element of the quotient set $\mathcal{P}^{\mathrm{cl}}/\simeq$. If γ° has a representative $\gamma = (x_0, \dots, x_l)$ for which $\gamma^- := (x_0, \dots, x_{l-1})$ is self-avoiding, then we call γ° an

(unbased) cycle. So, cycles can be seen as self-avoiding closed walks without a specified starting point. For any two representatives $\gamma_1, \gamma_2 \in \gamma^\circ$ it holds for their supports and traversed edges that $s(\gamma_1) = s(\gamma_2)$ and $e(\gamma_1) = e(\gamma_2)$. Hence, we can define the support of γ° by $s(\gamma^\circ) := s(\gamma_1)$ and the traversed edges of γ° by $e(\gamma^\circ) := e(\gamma_1)$.

The *occupation field* of a walk $\gamma = (x_0, x_1, \dots, x_l)$ is the map $\ell[\gamma] : \mathcal{X} \to \mathbb{N}_0$ defined by

$$\ell[\gamma](x) := \sum_{k=0}^{l} \mathbf{1}\{x_k = x\}, \tag{3.7}$$

where $\mathbf{1}$ denotes the indicator function. That is, $\ell[\gamma](x)$ denotes the *local time* spent by the walk γ at vertex x.

For each walk $\gamma = (x_0, x_1, \dots, x_l)$ we will define a self-avoiding walk called its *loop-erasure*. Iteratively define a sequence of self-avoiding walks $(\gamma_i)_{0 \le i \le l}$ as follows. Set $\gamma_0 := x_0$. For each $i \in [l]$, given $\gamma_{i-1} = (y_0, \dots, y_m)$ we define

$$\gamma_i := \begin{cases} (y_0, y_1, \dots, y_{m-1}, y_m, x_l) & \text{if } x_i \notin s(\gamma_{i-1}) \\ (y_0, y_1, \dots, y_{k_i - 1}, y_{k_i}) & \text{if } x_i \in s(\gamma_{i-1}), \end{cases} \tag{3.8}$$

where $k_i := \inf\{j \in [m]_0 \colon\ y_j = x_i\}$. The *loop-erasure* of γ is defined as

$$\mathrm{LE}[\gamma] := \gamma_l. \tag{3.9}$$

3.1.2.2 Wilson's algorithm using loop-erased random walks

Equip $\mathcal{X}$ with an arbitrary ordering, $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$. Let $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ be independent copies of the random walk X, with each $X^{(i)}$ starting from vertex x_i. Let $T^{(1)}, T^{(2)}, \dots, T^{(n)}$ be i.i.d. $\mathbb{N}_0$-valued[1] geometric killing times with success parameter $\frac{q}{q+\delta}$, that is $\mathbb{P}(T^{(i)} = k) = \left(\frac{\delta}{q+\delta}\right)^k \frac{q}{q+\delta}$. Iteratively define a set of vertices $V^{(i)}$ and a loop-erased random walk $\Gamma^{(i)}$ as follows. Set $V^{(0)} := \varnothing$. For each $i \in [n]$ write

$$\tau^{(i)} := T^{(i)} \wedge \inf\{k \in \mathbb{N}_0 \colon\ X_k^{(i)} \in V^{(i-1)}\} \tag{3.10}$$

to denote the minimum of $T^{(i)}$ and the first hitting time of $V^{(i-1)}$ by $X^{(i)}$, and define

$$\Gamma^{(i)} := \mathrm{LE}\left[(X_k^{(i)})_{0 \le k \le \tau^{(i)}}\right], \text{ and } V^{(i)} := V^{(i-1)} \cup s(\Gamma^{(i)}). \tag{3.11}$$

Theorem 3.1 (Wilson [77]). *The set of edges*

$$\bigcup_{i \in [n]} e(\Gamma^{(i)}) \tag{3.12}$$

is a Kirchhoff forest of intensity q.

While theorem 3.1 shows that the law of the rooted forest obtained by Wilson's procedure does not depend on the chosen vertex ordering, conditionally on the random walks $X^{(1)}, X^{(2)}, \dots, X^n$ the realization of the rooted forest does.

[1] In this chapter we adopt the convention that all geometric random variables are supported on $\mathbb{N}_0$.

3.1.2.3 Diaconis-Fulton stack representation

Diaconis and Fulton introduced an alternative method to index the randomness of a Markov chain [21]. Rather than using time as the index variable, they showed how a Markov chain can be constructed from space indexed randomness. It will be useful to use their representation for the random walks employed in Wilson's procedure, as doing so provides us with an alternative perspective on the entire procedure.

For each $x \in \mathcal{X}$ let $(A_i(x))_{i\in\mathbb{N}_0}$ be an independent sequence of i.i.d. $\mathcal{X}$-valued random variables with law

$$\mathbb{P}(A_i(x) = y) := \tfrac{1}{\delta} w(x, y), \quad \text{for all } y \in \mathcal{X}. \tag{3.13}$$

The random walk X can be constructed from the collection $\{(A_i(x))_{i\in\mathbb{N}_0} \colon\ x \in \mathcal{X}\}$, by iteratively setting

$$X_{k+1} := A_{i_k}(X_k), \quad \text{with } i_k := \ell[(X_j)_{0\le j\le k-1}](X_k), \tag{3.14}$$

where we use the convention that $\ell[\varnothing] := \underline{0}$, so that $i_0 = 0$.

One can imagine that an infinite stack of arrows $(A_i(x))_{i\in\mathbb{N}_0}$ is attached underneath each vertex x, with arrow $A_i(x)$ directly on top of arrow $A_{i+1}(x)$. Whenever the random walk visits a vertex, it reads the current top arrow from the stack of that vertex to determine its next step, after which that arrow is deleted from the stack.

We further introduce a collection $\{(B_i(x))_{i\in\mathbb{N}_0} \colon\ x \in \mathcal{X}\}$, where all $B_i(x)$ are i.i.d $\mathrm{Ber}(\frac{q}{q+\delta})$ random variables independent of the arrows. The variable $B_i(x)$ should be interpreted as the random color of arrow $A_i(x)$, where outcome 1 represents that the arrow is red and outcome 0 represents that the arrow is green. These colors are used to define the geometric killing time of the random walk. Whenever the random walk reads a green arrow, it makes the jump indicated by the arrow. But if it reads a red arrow, then it is killed instead.

3.1.2.4 Wilson's algorithm using cycle popping

Equip the set of unbased cycles with an arbitrary well-ordering. Set $\mathbf{d}_0 := \underline{0}$, where $\underline{0} \in \mathbb{N}_0^{\mathcal{X}}$ denotes the all-0 vector. For $i \in \mathbb{N}_0$ we iteratively define a random set of edges $E_i := \{(x, A_{\mathbf{d}_i(x)}(x)) \in \mathcal{E} \colon\ x \in \mathcal{X},\ B_{\mathbf{d}_i(x)}(x) = 0\}$ and let $\mathcal{P}_i^{\mathrm{cl}} := \{\gamma \in \mathcal{P}^{\mathrm{cl}} \colon\ e(\gamma) \subseteq E_i\}$ denote the set of closed walks whose edges are contained in E_i. Define the random cycle Γ_i° to be the minimal element of $\mathcal{P}_i^{\mathrm{cl}}/\simeq$ whenever $\mathcal{P}_i^{\mathrm{cl}}$ is non-empty, and set $\Gamma_i^\circ := \varnothing$ otherwise. Then we define $\mathbf{d}_{i+1} := \mathbf{d}_i + \mathbf{1}_{s(\Gamma_i^\circ)}$, where $\mathbf{1}_{s(\Gamma_i^\circ)} \in \{0,1\}^{\mathcal{X}}$ denotes the indicator of the support of Γ_i°. That is, at each iteration i one cycle is deleted (or 'popped') from the top of the DF-stacks, and $\mathbf{d}_i(x)$ denotes the number of deleted arrows from the stack at vertex x until iteration i.

Write $i^* := \min\{i \in \mathbb{N}_0 \colon\ \Gamma_i^\circ = \varnothing\}$ to denote the time-step when the procedure terminates.

Theorem 3.2 (Wilson, [77]). *The set of edges E_{i^*} is a Kirchhoff forest of intensity q.*

While the cycle popping procedure makes use of an ordering to choose a cycle, the realization of the Kirchhoff forest does not depend on the chosen ordering. We denote the rooted forest E_{i^*} obtained from the stacks $\{(A_i(x), B_i(x))_{i\in\mathbb{N}_0} : \ x \in \mathcal{X}\}$ by

$$\text{CyclePopping}(\{(A_i(x), B_i(x))_{i\in\mathbb{N}_0} : \ x \in \mathcal{X}\}) := E_{i^*}. \tag{3.15}$$

3.2 Coupled forests and their occupation fields

The intensity parameter q of the Kirchhoff forest obtained by cycle popping the stacks $\{(A_i(x), B_i(x))_{i\in\mathbb{N}_0} : \ x \in \mathcal{X}\}$ is determined solely by the parameter of the Bernoulli distribution of the colors. Rather than giving each arrow a single color, we can give a dynamic color $(B_i^t(x))_{t\geq 0}$ to each arrow that changes as time progresses. We will use a common method to couple together Bernoulli variables with different parameters, by making them depend on a single uniform random variable.

Consider a collection $\{(U_i(x))_{i\in\mathbb{N}_0} : \ x \in \mathcal{X}\}$, where all $U_i(x) \sim \text{Unif}(0,1)$ are i.i.d random variables, independent of the arrows $\{(A_i(x))_{i\in\mathbb{N}_0} : \ x \in \mathcal{X}\}$. For each $x \in \mathcal{X}$, $i \in \mathbb{N}_0$ and $t \geq 0$ we define a Bernoulli random variable

$$B_i^t(x) := \mathbf{1}\{U_i(x) < \tfrac{1}{1+\delta t}\}. \tag{3.16}$$

Definition 3.2.1. The *coupled forest process* $(\Phi_{1/t})_{t\geq 0}$ is the $\mathcal{F}$-valued stochastic process that is obtained from $\{(A_i(x), U_i(x))_{i\in\mathbb{N}_0} : \ x \in \mathcal{X}\}$ by defining

$$\Phi_{1/t} := \text{CyclePopping}(\{(A_i(x), B_i^t(x))_{i\in\mathbb{N}_0} : \ x \in \mathcal{X}\}). \tag{3.17}$$

■

Theorem 3.3 (Avena & Gaudillière [6]). *The coupled forest process $(\Phi_{1/t})_{t\geq 0}$ has the following properties:*

(*i*) *for fixed $t > 0$ the marginal $\Phi_{1/t}$ is a Kirchhoff forest of intensity $1/t$;*

(*ii*) *the process $(\Phi_{1/t})_{t\geq 0}$ has càdlàg and piece-wise constant trajectories;*

(*iii*) *the process $(\Phi_{1/t})_{t\geq 0}$ satisfies the Markov property.*

The coupled forest process starts, at time $t = 0$, consisting of only isolated vertices. At random times, exactly one of the current roots 'wakes up'. When that happens either its tree coalesces onto another tree, or its tree fragments into smaller trees, that each have the possibility to coalesce onto other trees.

The coalescing dynamic is more prominent than the fragmentation, so that as time progresses in expectation the number of trees decreasses.

3.2.1 The occupation field process

Let Φ_q be a Kirchhoff forest of intensity q obtained using the cycle popping procedure. Using the notation of section 3.1.2.4, we define the *occupation field (of Wilson's algorithm)* $\ell[\Phi_q]$ as the $\mathbb{N}^{\mathcal{X}}$-valued random variable

$$\ell[\Phi_q](x) := \mathbf{d}_{i^*}(x) + 1. \tag{3.18}$$

If Φ_q is constructed using loop-erased random walks, i.e. $\Phi_q := \bigcup_{i\in[n]} e\left(\Gamma^{(i)}\right)$, then we equivalently have that

$$\ell[\Phi_q](x) = \mathbf{1}\{x \in \rho(\Phi_q)\} + \sum_{i=1}^{n} \ell[(X_k^{(i)})_{0\le k\le \tau^{(i)}-1}](x). \tag{3.19}$$

That is, the Wilson occupation field equals the sum of the occupation fields of the stopped random walks used in the procedure, where we only count the local time contribution of the final step of random walk $X^{(i)}$ if $X^{(i)}$ is killed before it hits any of the trajectories of the previous random walks.

The notation $\ell[\Phi_q]$ might incorrectly suggest that the occupation field is an observable of the Kirchhoff forest Φ_q. Therefore, we emphasize that the occupation field is constructed from a realization of Wilson's algorithm and not just from the resulting forest.

Although not explicitly mentioned as a result, in [77, proof of Thm 1] Wilson used the following observation.

Lemma 3.4 (Wilson [77], *proof of Thm. 1 therein*). *The Kirchhoff forest Φ_q is independent of the occupation field $\ell[\Phi_q]$.*

The construction of the coupled forest process, as given in definition 3.2.1, does not only couple together a family of Kirchhoff forests, but also their Wilson occupation fields. For notational brevity, we denote the occupation field of $\Phi_{1/t}$ by

$$N_t := \ell[\Phi_{1/t}], \tag{3.20}$$

thus defining an *occupation field process* $(N_t)_{t\ge 0}$.

The occupation field is closely related to the *running time* M_t of Wilson's algorithm. The latter being obtained by taking a sum of the occupation field over all vertices,

$$M_t := \sum_{x\in\mathcal{X}} N_t(x). \tag{3.21}$$

Equivalently, the running time equals the total number of edges traversed by the n random walks used in the construction plus the number of roots of the obtained forest. That is, using the notation of section 3.1.2.2

$$M_{1/q} = r(\Phi_q) - n + \sum_{i\in[n]} \tau^{(i)}. \tag{3.22}$$

In [61, prop. 1] Marchal expresses the probability generating function (pgf) of the Wilson running time in terms of the determinant of the transition matrix of the random walk used in Wilson's procedure. Here we apply this result to our specific choice of killed random walk with transition matrix $(I - \frac{1}{\delta}L)$ and killing rate $\frac{q}{q+\delta}$. We denote by $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$ the spectrum of the Laplacian L.

Proposition 3.5 (Marchal [61]). *For $z \in (0,1)$ it holds that*

$$\mathbb{E}\left[z^{M_{1/q}}\right] = \frac{\det\left[\frac{q}{q+\delta}I + \frac{1}{q+\delta}L\right] z^n}{\det\left[I - \left(I - \frac{q}{q+\delta}I - \frac{1}{q+\delta}L\right) z\right]} = \prod_{j<n} \frac{\frac{q+\lambda_j}{q+\delta} z}{1 - \left(1 - \frac{q+\lambda_j}{q+\delta}\right) z}. \tag{3.23}$$

If $\lambda_j \in \mathbb{R}$ and $\delta \geq \lambda_j$ for all j, then we recognize each of the n factors as the pgf of an $\mathbb{N}$-valued geometric random variable with success parameter $\frac{q+\lambda_j}{q+\delta}$. Hence, in this case we find that $M_{1/q}$ is the sum of n independent geometrics.

3.2.1.1 Different random walks

As mentioned in section 3.1.2, Wilson's algorithm has some freedom in the choice of the transition matrix of the employed random walks. While the choice of random walk does not affect the law of the resulting rooted forest, the distribution of the occupation field does depend on the choice of random walk. In particular, since the transition matrix $(I - \frac{1}{\delta}L)$ used in this chapter depends on a parameter δ, the occupation field will depend on the parameter δ as well.

It is possible to construct the coupled forest process in a parameter free manner, e.g. by employing the simple random walk with transition matrix $D^{-1}A$, where A denotes the (weighted) adjacency matrix of $\mathcal{G}$ and D is the diagonal matrix such that $L = D - A$. However, the results presented here treat the δ dependent occupation field process.

The choice for the transition matrix $(I - \frac{1}{\delta}L)$ has several advantages. Firstly, the killing time of the random walk can be taken independent of the trajectory of the random walk, which simplifies some of the computations. More importantly, the spectrum of the transition matrix is obtained directly from the spectrum of the Laplacian L, which is relevant since several observables of the occupation field can be expressed in terms of the spectrum of the transition matrix, see e.g. proposition 3.5 above. Hence, using our choice of transition matrix ensures that these observables are expressible in terms of the Laplacian spectrum $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$. This enables the possibility of future work to develop Laplacian spectrum estimation procedures based on observables of the occupation field process. However, development of such procedures lies outside of the scope of the present chapter.

3.3 Result: Law of the occupation field process

The result of this chapter is a complete description of the law of the occupation field process $(N_t)_{t\geq 0}$, of which theorem 3.6 and corollary 3.6.1 give two equivalent formulations.

For stating the result it will be convenient to consider three distinct probability spaces. On one probability space we define the DF-stacks $\{(A_i(x), U_i(x))_{i\in\mathbb{N}_0} : x \in \mathcal{X}\}$, consisting of arrows and uniform killing marks, and all random variables coupled thereto, e.g. the occupation field process $(N_t)_{t\geq 0}$. The law of the DF-stacks we denote by $\mathbb{P}$. We will further consider, on a second probability space, the killed random walk $(X_k)_{0\leq k\leq T_t}$, which is killed at an $\mathbb{N}_0$-supported geometric time $T_t \sim \mathrm{Geom}_{\mathbb{N}_0}\left(\frac{1}{1+\delta t}\right)$ that is independent of X, for some parameter $t \geq 0$. Denote by $\mathbf{P}_{\mathrm{unif}}$ the joint law of X and T_t, with X starting from a uniformly chosen vertex, and write $\mathbf{P}_x(\cdot) := \mathbf{P}_{\mathrm{unif}}(\cdot \mid X_0 = x)$. Any other auxiliary random variables, that are coupled to neither the DF-stacks nor to the killed random walk, will be defined on a third space

with law P and accompanying expectation E.

We define the matrix $K_{1/t} := \frac{1}{t}(\frac{1}{t}I + L)^{-1}$, and let $\mathrm{Tr}[\cdot]$ denote the trace operator.

Theorem 3.6. *The occupation field process $(N_t)_{t\geq 0}$ is a Markov process with piecewise constant càdlàg trajectories and distribution*

$$(N_t)_{t\geq 0} \stackrel{d}{=} \left(\underline{1} + \int_0^t \Lambda_s \ \Psi(\mathrm{d}s)\right)_{t\geq 0}, \tag{3.24}$$

where for all $t > 0$ the variables Λ_t are independent $\mathbb{N}_0^{\mathcal{X}}$-valued random variables with law

$$P(\Lambda_t = \mathbf{m}) := \mathbf{P}_{\mathrm{unif}}(\ell[(X_k)_{0\leq k\leq T_t}] = \mathbf{m} \mid X_{T_t+1} = X_0), \quad \textit{for all } \mathbf{m} \in \mathbb{N}_0^{\mathcal{X}}, \tag{3.25}$$

and Ψ is an inhomogeneous Poisson point process on $(0,\infty)$ with intensity measure

$$\mu((a,b]) := \int_a^b \mathrm{Tr}\left[\tfrac{1}{t}(K_{1/t} - \tfrac{1}{1+\delta t}I)\right] \ \mathrm{d}t, \tag{3.26}$$

that is independent of all Λ_t.

According to theorem 3.6, at time $t = 0$ the occupation field N_0 equals 1 at each vertex. As t increases the jump-times of the occupation field process are a Poisson point process with explicit rate

$$\kappa_t := \mathrm{Tr}\left[\tfrac{1}{t}(K_{1/t} - \tfrac{1}{1+\delta t}I)\right]. \tag{3.27}$$

At each of its jump times the occupation field is increased by a random amount whose distribution is that of the occupation field of a killed random walk that is conditioned to make at least one step and to be killed at its uniformly chosen starting point.

Conditional on the event $T_t \geq 1$ the distribution of T_t is the same as the unconditional distribution of $T_t + 1$. This fact is used in theorem 3.6 to simplify notation.

3.3.0.1 Closed walk decomposition

In [55] Le Jan showed that the occupation field of Wilson's algorithm has the same distribution as the occupation field of a Poissonian loop-ensemble. The statement of corollary 3.6.1 is a rephrasing of theorem 3.6, that clarifies how the connection between Wilson's occupation field and the loop-ensemble occupation field extends to our dynamical setting.

Define the random set of *jump times* of the occupation field process by

$$\mathcal{T} := \{t \in (0,\infty) \colon \quad N_t \neq \lim_{s\uparrow t} N_s\}. \tag{3.28}$$

So, $\mathcal{T}$ is distributed as the support of the Poisson point process Ψ.

Rather then seeing Ψ as a single Poisson point process, we can decompose Ψ into multiple Poisson point processes that are associated to the closed walks in $\mathcal{G}$.

Conditionally on these Poisson processes the increment of the occupation field at a jump time τ can then be determined deterministically by observing which of these Poisson point processes contains τ.

For a closed walk $\gamma = (x_0, x_1, \ldots, x_l) \in \mathcal{P}^{\mathrm{cl}}$ write $\gamma^- := (x_0, x_1, \ldots x_{l-1})$ to denote the walk obtained by omitting the last step from γ.

Corollary 3.6.1. *For all closed walks $\gamma \in \mathcal{P}^{\mathrm{cl}}$ let Ψ_γ be independent Poisson point processes with respective intensity measures*

$$\mu_\gamma((a,b]) := w(e(\gamma)) \int_a^b \frac{1}{(1+\delta t)^2} \left(\frac{t}{1+\delta t}\right)^{l-1} \mathrm{d}t. \tag{3.29}$$

Then it holds that

$$(N_t)_{t\geq 0} \overset{d}{=} \left(\underline{1} + \sum_{\gamma \in \mathcal{P}^{\mathrm{cl}}} \ell[\gamma^-]\ \Psi_\gamma((0,t]) \right)_{t\geq 0}. \tag{3.30}$$

For fixed $t > 0$ we can define a measure ν_t on closed walks by

$$\nu_t(\gamma) := \mu_\gamma([0,t)) = w(e(\gamma)) \int_0^t \frac{1}{(1+\delta s)^2} \left(\frac{s}{1+\delta s}\right)^{l-1} \mathrm{d}s = \frac{1}{l} \frac{w(e(\gamma))}{(1/t+\delta)^l}. \tag{3.31}$$

The measure ν_t indeed corresponds to intensity measure of the Poissonian loop-ensemble, if the loops in the ensemble are obtained from the random walk with transition matrix $(I - \frac{1}{\delta}L)$ with homogeneous killing rate $\frac{1}{1+\delta t}$, see Le Jan [56, eq (2.4)].

Hence, corollary 3.6.1 shows that the 'loop measure' ν_t can be constructed by taking an integral over different random walk killing rates. This integration further explains the appearance of the factor $\frac{1}{l}$, which for continuous-time random walk ensembles causes the loop measure to explode as the length of the loop approaches 0.

3.4 Proofs

The proof of theorem 3.6 is divided into three parts, which are each subdivided into various lemmas. In the first part it is shown that the occupation field process has independent increments. This fact is then used in the second part to compute the distribution of the random set of jump times of the process. In the third part we compute the occupation field increment distribution at jump times.

3.4.1 Independent increments

The argument that is employed in [6] to demonstrate the Markovianity of the coupled forest process $(\Phi_{1/t})_{t\geq 0}$, also shows that the joint process $(\Phi_{1/t}, N_t)_{t\geq 0}$ is Markovian. More specifically, it shows that, for fixed $0 < s < t$, the distribution of $(\Phi_{1/t},\ N_t - N_s)$ depends on the joint history $(\Phi_{1/r}, N_r)_{0\leq r\leq s}$ only through $\Phi_{1/s}$.

That is, for all $k \in \mathbb{N}$, $0 \leq t_1 < t_2 < \ldots < t_k = s$, $F, F_1, F_2, \ldots, F_k \in \mathcal{F}$, $\mathbf{n} \in \mathbb{N}_0^{\mathcal{X}}$, and all $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_k \in \mathbb{N}^{\mathcal{X}}$ with $\mathbb{P}((\Phi_{1/t_i}, N_{t_i})_{i\in[k]} = (F_i, \mathbf{m}_i)_{i\in[k]}) > 0$, it holds

that

$$\begin{aligned}\mathbb{P}(\Phi_{1/t} = F,\ N_t - N_s = \mathbf{n} \mid (\Phi_{1/t_i}, N_{t_i})_{i\in[k]} = (F_i, \mathbf{m}_i)_{i\in[k]}) \\ = \mathbb{P}(\Phi_{1/t} = F,\ N_t - N_s = \mathbf{n} \mid \Phi_{1/t_k} = F_k).\end{aligned} \tag{3.32}$$

The following lemma shows that the occupation field process $(N_t)_{t\geq 0}$ is itself Markovian as well.

Lemma 3.7. *Both the Wilson occupation field process $(N_t)_{t\geq 0}$ and the running time process $(M_t)_{t\geq 0}$ have independent increments.*

Proof. We will only prove the result for the occupation field process, as the proof of for the running time is analogous. Fix $0 < s < t$. We have to show that the increment of the occupation field $N_t - N_s$ is independent of $(N_r)_{0\leq r\leq s}$.

Consider the joint process $(\Phi_{1/t}, N_t)_{t\geq 0}$, which is constructed from the DF-stacks $\{(A_i(x), U_i(x))_{i\in\mathbb{N}_0}\colon\ x \in \mathcal{X}\}$. In the first part of the proof we show that $\Phi_{1/t}$ is independent of $(N_s)_{0\leq s\leq t}$, for which it is sufficient that the events $\{\Phi_{1/t} = F\}$ and $\bigcap_{i\in[k]}\{N_{t_i} \succeq \mathbf{m}_i + \underline{1}\}$ are independent[2], for arbitrary $k \in \mathbb{N}$, $0 < t_1 < t_2 < \ldots < t_k \leq t$, $F \in \mathcal{F}$ and $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_k \in \mathbb{N}_0^{\mathcal{X}}$. Here $\succeq$ denotes the natural partial order on $\mathbb{N}_0^{\mathcal{X}}$ given by $\mathbf{n} \preceq \mathbf{m}$ if $\mathbf{n}(x) \leq \mathbf{m}(x)$ for all $x \in \mathcal{X}$.

Fix $\mathbf{m} \in \mathbb{N}_0^{\mathcal{X}}$ such that $\mathbf{m} \succeq \mathbf{m}_k$. There exists some subset

$$\mathcal{S}_F \subseteq \prod_{x\in\mathcal{X}} (\{y \in \mathcal{X}\colon\ (x, y) \in \mathcal{E}\} \times [0, 1])$$

such that[3]

$$\{N_t = \mathbf{m} + \underline{1}, \Phi_{1/t} = F\} = \{N_t = \mathbf{m} + \underline{1},\ \{(A_{\mathbf{m}(x)}(x), U_{\mathbf{m}(x)}(x))\colon\ x \in \mathcal{X}\} \in \mathcal{S}_F\}.$$

So, conditionally on the event $\{N_t = \mathbf{m} + \underline{1}\}$, the event $\{\Phi_{1/t} = F\}$ depends only on the variables $\{(A_{\mathbf{m}(x)}(x), U_{\mathbf{m}(x)}(x))\colon\ x \in \mathcal{X}\}$.

For each $i \in [k]$ whether the cycle popping procedure deletes the arrow above layer $\mathbf{m}_i$ does not depend on the arrows in layer $\mathbf{m}_i$ and the arrows below that layer, i.e. the event $\{N_{t_i} \succeq \mathbf{m}_i + \underline{1}\}$ is determined by the variables $\{(A_{i_x}(x), U_{i_x}(x))\colon\ x \in \mathcal{X},\ i_x = 1, 2, \ldots, \mathbf{m}_i(x) - 1\}$. Hence, the intersection $\bigcap_{i\in[k]}\{N_{t_i} \succeq \mathbf{m}_i + \underline{1}\}$ is determined by the variables $\{(A_{i_x}(x), U_{i_x}(x))\colon\ x \in \mathcal{X},\ i_x = 1, 2, \ldots, \mathbf{m}_k(x) - 1\}$. Since $\mathbf{m} \succeq \mathbf{m}_k$, we have in particular that this intersection does not depend on the variables $\{(A_{\mathbf{m}(x)}(x), U_{\mathbf{m}(x)}(x))\colon\ x \in \mathcal{X}\}$.

As this holds for all $\mathbf{m}$, it follows that the events $\bigcap_{i\in[k]}\{N_{t_i} \succeq \mathbf{m}_i + \underline{1}\}$ and $\{\Phi_{1/t} = F\}$ are conditionally independent given N_t. By lemma 3.4, the variables $\Phi_{1/t}$

[2]The addition of the all-ones vector $\underline{1}$ is an administrative trick, that is required since $N_0 = \underline{1}$, while the indices of the DF-stacks start at 0.

[3]Explicitly the subset $\mathcal{S}_F$ is given by

$$\mathcal{S}_F = \left(\prod_{x\in\mathcal{X}\setminus\rho(F)} \{y \in \mathcal{X}\colon\ (x, y) \in F\} \times [\tfrac{1}{1+\delta t}, 1]\right) \times \left(\prod_{x\in\rho(F)} \{y \in \mathcal{X}\colon\ (x, y) \in \mathcal{E}\} \times [0, \tfrac{1}{1+\delta t})\right).$$

and N_t are independent. It follows that $\bigcap_{i\in[k]}\{N_{t_i} \succeq \mathbf{m}_i + \underline{1}\}$ and $\{\Phi_{1/t} = F\}$ are also unconditionally independent, which completes the first part of the proof.

It follows from eq. (3.32) for all $\mathbf{n} \in \mathbb{N}_0^{\mathcal{X}}$ that

$$\mathbb{P}(N_t - N_s = \mathbf{n} \mid \Phi_{1/s},\ (N_r)_{0\le r\le s}) = \mathbb{P}(N_t - N_s = \mathbf{n} \mid \Phi_{1/s}). \tag{3.33}$$

Therefore, the increment $N_t - N_s$ is conditionally independent of $(N_r)_{0\le r\le s}$ given $\Phi_{1/s}$. By the first part of the proof $(N_r)_{0\le r\le s}$ and $\Phi_{1/s}$ are independent, from which it follows that $N_t - N_s$ and $(N_r)_{0\le r\le s}$ are also unconditionally independent. □

3.4.2 Characterization of jump times

Lemma 3.8 (Characterization of jump times). *The random atomic measure $\sum_{t\in\mathcal{T}}\delta_t$ supported on the set $\mathcal{T}$ of jump times of the occupation field process is a Poisson point process on $(0,\infty)$ with intensity measure*

$$\mu((a,b]) := \int_a^b \operatorname{Tr}\left[\tfrac{1}{t}(K_{1/t} - \tfrac{1}{1+\delta t}I)\right]\,\mathrm{d}t. \tag{3.34}$$

Proof. By lemma 3.7 and Kingman's representation theorem [46], the random measure $\sum_{t\in\mathcal{T}}\delta_t$ is a Poisson point process on $(0,\infty)$.

Note that the jump times $\mathcal{T}$ of the occupation field process $(N_t)_{t\ge0}$ are equal to the jump times of the running time process $(M_t)_{t\ge0}$.

Using proposition 3.5 and lemma 3.7 we find that for fixed $0 < s < t$ the pgf of $M_t - M_s$ is given by

$$\begin{aligned}\mathbb{E}\left[z^{M_t - M_s}\right] &= \prod_{j<n}\frac{\frac{1+\lambda_j t}{1+\delta t}\left(1-\left(1-\frac{1+\lambda_j s}{1+\delta s}\right)z\right)}{\frac{1+\lambda_j s}{1+\delta s}\left(1-\left(1-\frac{1+\lambda_j t}{1+\delta t}\right)z\right)}\\ &= \prod_{j<n}\frac{(1+\lambda_j t)\,(1+\delta s-(\delta-\lambda_j)\,sz)}{(1+\lambda_j s)\,(1+\delta t-(\delta-\lambda_j)\,tz)}.\end{aligned} \tag{3.35}$$

Its intensity measure follows from eq. (3.35), since evaluating the pgf of $M_t - M_s$ at $z = 0$ gives us that

$$e^{\mu([s,t))} = \mathbb{P}(M_t = M_s) = \prod_{j<n}\frac{(1+\lambda_j t)\,(1+\delta s)}{(1+\lambda_j s)\,(1+\delta t)}. \tag{3.36}$$

By taking the logarithm we find that

$$\begin{aligned}\mu([a,b)) &= \sum_{j<n}\log\left(\frac{(1+\lambda_j a)(1+\delta b)}{(1+\delta a)(1+\lambda_j b)}\right) = \int_a^b\sum_{j<n}\frac{\delta-\lambda_j}{(1+\delta t)(1+\lambda_j t)}\,\mathrm{d}t\\ &= \int_a^b \operatorname{Tr}\left[\tfrac{1}{t}(K_{1/t} - \tfrac{1}{1+\delta t}I)\right]\,\mathrm{d}t.\end{aligned}$$

□

3.4.3 Distribution of increments at jump times

The remainder of the proof of theorem 3.6 consists of identifying the law of the increments at the jump times.

For this purpose we consider an extended graph $\mathcal{G}^\star = (\mathcal{X}^\star, \mathcal{E}^\star, w^\star)$, that depends on a parameter $\mathbf{z} \in (0,1)^{\mathcal{X}}$. The graph $\mathcal{G}^\star$ has vertices $\mathcal{X}^\star := \mathcal{X} \cup \{\bigstar, \blacklozenge\}$, directed edges $\mathcal{E}^\star := \mathcal{E} \cup \{(x, \bigstar)\colon\ x \in \mathcal{X} \cup \{\bigstar\}\} \cup \{(x, \blacklozenge)\colon\ x \in \mathcal{X} \cup \{\blacklozenge\}\}$, and edge weights

$$w^\star(x,y) := \begin{cases} w(x,y) & \text{if } x,y \in \mathcal{X}, \\ \frac{1}{t}\ \frac{1-\mathbf{z}(x)}{\mathbf{z}(x)} & \text{if } x \in \mathcal{X},\ y = \bigstar, \\ \delta\ \frac{1-\mathbf{z}(x)}{\mathbf{z}(x)} & \text{if } x \in \mathcal{X},\ y = \blacklozenge, \\ 0 & \text{otherwise.} \end{cases} \tag{3.37}$$

That is, $\mathcal{G}^\star$ is obtained from $\mathcal{G}$ by adding two absorbing vertices, and adding edges from all vertices in $\mathcal{X}$ pointing to these new vertices. The parameter $\mathbf{z}$ regulates the edge weights of the added edges. We denote the graph Laplacian of $\mathcal{G}^\star$ by $L^\star$.

In lemma 3.9 below, we will compute the probability generating function (pgf) of the Wilson occupation field at a fixed intensity.

Lemma 3.9 (PGF of Wilson occupation field). *Fix $t > 0$. For $\mathbf{z} \in (0,1)^{\mathcal{X}}$ the Wilson occupation field has multi-dimensional probability generating function*

$$\mathbb{E}\left[\prod_{x\in\mathcal{X}} \mathbf{z}(x)^{N_t(x)}\right] = \frac{\det\left[\frac{1}{t}I + L\right]}{\det\left[\left(\frac{1}{t}+\delta\right)\operatorname{diag}\left(\frac{1-\mathbf{z}}{\mathbf{z}}\right) + \frac{1}{t}I + L\right]}, \tag{3.38}$$

where $\frac{1-\mathbf{z}}{\mathbf{z}}(x) := \frac{1-\mathbf{z}(x)}{\mathbf{z}(x)}$ for all $x \in \mathcal{X}$.

Proof. The multi-dimensional pgf can be expressed as

$$\mathbb{E}\left[\prod_{x\in\mathcal{X}} \mathbf{z}(x)^{N_t(x)}\right] = \mathbb{P}\left(\bigcap_{x\in\mathcal{X}} \{N_t(x) \le T(x)\}\right), \tag{3.39}$$

where $\{T(x)\colon\ x \in \mathcal{X}\}$ is any arbitrary collection of independent random variables with distribution $T(x) \sim \mathrm{Geom}_{\mathbb{N}_0}(1 - \mathbf{z}(x))$. We will construct a specific such collection of geometrics, for which the events $\{N_t(x) \le T(x)\}$ have an interpretation in the context of Wilson's algorithm that allows us to evaluate the right hand side of eq. (3.39).

Consider the extended graph $\mathcal{G}^\star$ defined above in eq. (3.37). We construct a collection of DF-stacks $\{(A_i^\star(x), B_i^\star(x))_{i\in\mathbb{N}_0}\colon\ x \in \mathcal{X}^\star\}$ on the extended graph $\mathcal{G}^\star$, which will be coupled to the DF-stacks $\{(A_i(x), B_i(x))_{i\in\mathbb{N}_0}\colon\ x \in \mathcal{X}\}$ of the original graph $\mathcal{G}$. For for all $i \in \mathbb{N}_0$ we set the arrows by $A_i^\star(\bigstar) = \bigstar$, $A_i^\star(\blacklozenge) = \blacklozenge$ and independently set

$$A_i^\star(x) = \begin{cases} A_i(x) & \text{with probability } \frac{1/t+\delta}{1/t+\delta+w^\star(x,\bigstar)+w^\star(x,\blacklozenge)} \\ \bigstar & \text{with probability } \frac{w^\star(x,\bigstar)}{1/t+\delta+w^\star(x,\bigstar)+w^\star(x,\blacklozenge)} \\ \blacklozenge & \text{with probability } \frac{w^\star(x,\blacklozenge)}{1/t+\delta+w^\star(x,\bigstar)+w^\star(x,\blacklozenge)} \end{cases} \quad \text{for all } x \in \mathcal{X}. \tag{3.40}$$

The killing marks are defined by $B_i^\star(\bigstar) = 1$, $B_i^\star(\blacklozenge) = 1$ and

$$B_i^\star(x) = \begin{cases} B_i(x) & \text{if } A_i^\star(x) = A_i(x) \\ 0 & \text{if } A_i^\star(x) \in \{\bigstar, \blacklozenge\} \end{cases} \quad \text{for all } x \in \mathcal{X}. \tag{3.41}$$

These DF-stacks do not have the distribution given in eq. (3.13), and as a consequence the killed Markov chain on $\mathcal{G}^\star$ constructed from these stacks, as defined in section 3.1.2.3, does not have transition matrix $I - \frac{1}{\delta^\star}L^\star$ for some parameter $\delta^\star$. However, the law of the loop-erased trajectory of the killed Markov chain constructed from these stacks, is identical to the law of the loop-erased trajectory of the killed random walk with transition matrix $I - \frac{1}{\delta^\star}L^\star$. Hence, applying the cycle popping procedure to the stacks $\{(A_i^\star(x), B_i^\star(x))_{i\in\mathbb{N}_0}\colon\ x \in \mathcal{X}^\star\}$ does produce a Kirchhoff forest of the graph $\mathcal{G}^\star$ of intensity $1/t$, which we denote by $\Phi^\star_{1/t}$.

Cycle popping the DF-stacks on $\mathcal{G}^\star$ also produces a field of stack depths $\ell^\star[\Phi^\star_{1/t}]$, as defined in eq. (3.18), where $\ell[\Phi^\star_{1/t}](x)-1$ denotes the index of the arrow at x that is contained in $\Phi^\star_{1/t}$. We remark that since the stacks on $\mathcal{G}^\star$ define a random walk with a different transition matrix, the field $\ell^\star[\Phi^\star_{1/t}]$ is not distributed as a Wilson occupation field on $\mathcal{G}^\star$, as expained in section 3.2.1.1.

Define the collection of geometrics $\{T(x)\colon\ x \in \mathcal{X}\}$ by

$$T(x) := \min\{i \in \mathbb{N}_0\colon\ A_i^\star(x) \in \{\bigstar, \blacklozenge\}\}, \qquad \text{for all } x \in \mathcal{X}, \tag{3.42}$$

which indeed has success parameter $\frac{w^\star(x,\bigstar)+w^\star(x,\blacklozenge)}{1/t+\delta+w^\star(x,\bigstar)+w^\star(x,\blacklozenge)} = 1 - \mathbf{z}(x)$, and is independent of $\{T(y)\colon\ y \in \mathcal{X}\setminus\{x\}\}$ and $\ell[\Phi_{1/t}]$.

Since $T(x)$ denotes the index of the topmost arrow that points to one of the two added vertices, it holds that

$$\{\ell[\Phi^\star_{1/t}](x) \le T(x)\} = \{(x,\bigstar) \notin \Phi^\star_{1/t} \text{ and } (x,\blacklozenge) \notin \Phi^\star_{1/t}\}, \tag{3.43}$$

where we use that arrows of the form $(x,\bigstar)$ or $(x,\blacklozenge)$ cannot be part of a cycle in $\mathcal{G}^\star$, so they can never be popped by the cycle popping procedure. It follows that

$$\bigcap_{x\in\mathcal{X}} \{\ell[\Phi^\star_{1/t}](x) \le T(x)\} = \{\bigstar \text{ and } \blacklozenge \text{ are isolated vertices of } \Phi^\star_{1/t}\}. \tag{3.44}$$

By the coupling of $\{(A_i^\star(x), B_i^\star(x))_{i\in\mathbb{N}_0}\colon\ x \in \mathcal{X}^\star\}$ and $\{(A_i(x), B_i(x))_{i\in\mathbb{N}_0}\colon\ x \in \mathcal{X}\}$ it holds that

$$\bigcap_{x\in\mathcal{X}} \{\ell[\Phi^\star_{1/t}](x) \le T(x)\} = \bigcap_{x\in\mathcal{X}} \{\ell[\Phi_{1/t}](x) \le T(x)\}. \tag{3.45}$$

Hence, recalling that $\mathcal{F}$ denotes the set of rooted forests of $\mathcal{G}$, we have by the matrix-tree theorem that

$$\begin{aligned}
\mathbb{E}\left[\prod_{x\in\mathcal{X}} \mathbf{z}(x)^{N_t(x)}\right] &= \mathbb{P}(\bigstar \text{ and } \blacklozenge \text{ are isolated vertices of } \Phi^\star_{1/t}) \\
&= \mathbb{P}(\Phi^\star_{1/t} \in \mathcal{F}) = \frac{\frac{1}{t^2}\det[\frac{1}{t}I + L]}{\det[\frac{1}{t}I + L^\star]} \\
&= \frac{\det\left[\frac{1}{t}I + L\right]}{\det\left[\left(\frac{1}{t}+\delta\right)\operatorname{diag}\left(\frac{1-\mathbf{z}}{\mathbf{z}}\right) + \frac{1}{t}I + L\right]}.
\end{aligned}$$

□

Lemma 3.10 (PGF of increment of Wilson occupation field). *Fix $\mathbf{z} \in (0,1)^{\mathcal{X}}$. For the pgf of the increment $N_t - N_s$ of the Wilson occupation field conditioned on being non-zero, it holds in the limit $s \uparrow t$ that*

$$\mathbb{E}\left[\prod_{x\in\mathcal{X}} \mathbf{z}(x)^{N_t(x)-N_s(x)} \;\middle|\; N_t - N_s \neq \underline{0}\right] = \frac{\mathrm{Tr}\left[K^{\delta\mathbf{z}}_{t\mathbf{z},t}\right] - \frac{n}{1+\delta t}}{\mathrm{Tr}\left[K_{1/t}\right] - \frac{n}{1+\delta t}} + \mathcal{O}(t-s), \tag{3.46}$$

where $K_{1/t} := \frac{1}{t}(\frac{1}{t}I + L)^{-1}$ and

$$K^{\delta\mathbf{z}}_{t\mathbf{z},t} := \left(\left(\tfrac{1}{t}+\delta\right)\mathrm{diag}\left(\tfrac{\underline{1}-\mathbf{z}}{\mathbf{z}}\right) + \tfrac{1}{t}I + L\right)^{-1}\left(\tfrac{1}{t}\mathrm{diag}\left(\tfrac{\underline{1}-\mathbf{z}}{\mathbf{z}}\right) + \tfrac{1}{t}I\right). \tag{3.47}$$

Proof. Write

$$\mathbb{E}\left[\prod_{x\in\mathcal{X}} \mathbf{z}(x)^{N_t(x)-N_s(x)} \;\middle|\; N_t - N_s \neq \underline{0}\right] = \frac{\mathbb{E}\left[\prod_{x\in\mathcal{X}} \mathbf{z}(x)^{N_t-N_s}\right] - \mathbb{P}(N_t - N_s = \underline{0})}{\mathbb{P}(N_t - N_s \neq \underline{0})}. \tag{3.48}$$

It holds by lemmas 3.7 and 3.9 that

$$\begin{aligned}\mathbb{E}\left[\prod_{x\in\mathcal{X}} \mathbf{z}(x)^{N_t-N_s}\right] &= \frac{\mathbb{E}\left[\prod_{x\in\mathcal{X}} \mathbf{z}(x)^{N_t}\right]}{\mathbb{E}\left[\prod_{x\in\mathcal{X}} \mathbf{z}(x)^{N_s}\right]} \\ &= \frac{\det\left[\frac{1}{t}I + L\right]\det\left[(1/s+\delta)\mathrm{diag}\left(\frac{\underline{1}-\mathbf{z}}{\mathbf{z}}\right) + \frac{1}{s}I + L\right]}{\det\left[\frac{1}{s}I + L\right]\det\left[(1/t+\delta)\mathrm{diag}\left(\frac{\underline{1}-\mathbf{z}}{\mathbf{z}}\right) + \frac{1}{t}I + L\right]}.\end{aligned}$$

As $s \uparrow t$, Jacobi's formula gives us that

$$\begin{aligned}\det\left[\tfrac{1}{s}I + L\right] &= \det\left[\tfrac{1}{t}I + L\right] - (t-s)\det\left[\tfrac{1}{t}I + L\right]\mathrm{Tr}\left[(\tfrac{1}{t}I + L)^{-1}(-\tfrac{1}{t^2}I)\right] + \mathcal{O}\left((t-s)^2\right) \\ &= \det\left[\tfrac{1}{t}I + L\right]\left(1 + \tfrac{(t-s)}{t}\mathrm{Tr}\left[K_{1/t}\right]\right) + \mathcal{O}\left((t-s)^2\right),\end{aligned}$$

and similarly that

$$\begin{aligned}&\det\left[(1/s+\delta)\mathrm{diag}\left(\tfrac{\underline{1}-\mathbf{z}}{\mathbf{z}}\right) + \tfrac{1}{s}I + L\right] \\ &= \det\left[(1/t+\delta)\mathrm{diag}\left(\tfrac{\underline{1}-\mathbf{z}}{\mathbf{z}}\right) + \tfrac{1}{t}I + L\right]\left(1 + \tfrac{(t-s)}{t}\mathrm{Tr}\left[K^{\delta\mathbf{z}}_{t\mathbf{z},t}\right]\right) + \mathcal{O}\left((t-s)^2\right).\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{E}\left[\prod_{x\in\mathcal{X}} \mathbf{z}(x)^{N_t(x)-N_s(x)}\right] &= \frac{1 + \frac{(t-s)}{t}\mathrm{Tr}\left[K^{\delta\mathbf{z}}_{t\mathbf{z},t}\right]}{1 + \frac{(t-s)}{t}\mathrm{Tr}\left[K_{1/t}\right]} + \mathcal{O}\left((t-s)^2\right) \\ &= 1 + \tfrac{(t-s)}{t}\mathrm{Tr}\left[K^{\delta\mathbf{z}}_{t\mathbf{z},t} - K_{1/t}\right] + \mathcal{O}\left((t-s)^2\right).\end{aligned}$$

From lemma 3.8 we know that

$$\begin{aligned}\mathbb{P}(N_t - N_s \neq \underline{0}) &= 1 - \prod_{j<n}\frac{(1+\delta s)(1+\lambda_j t)}{(1+\lambda_j s)(1+\delta t)} \\ &= (t-s)\sum_{j<n}\frac{\delta - \lambda_j}{(1+\lambda_j t)(1+\delta t)} + \mathcal{O}((t-s)^2).\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\prod_{x\in\mathcal{X}}\mathbf{z}(x)^{N_t-N_s}\;\middle|\; N_t-N_s\neq\underline{0}\right]\\
&=\frac{\frac{1}{t}\mathrm{Tr}\left[K^{\delta\mathbf{z}}_{t\mathbf{z},t}-K_{1/t}\right]+\sum_{j<n}\frac{\delta-\lambda_j}{(1+\lambda_j t)(1+\delta t)}}{\sum_{j<n}\frac{\delta-\lambda_j}{(1+\lambda_j t)(1+\delta t)}}+\mathcal{O}(t-s)\\
&=\frac{\mathrm{Tr}\left[K^{\delta\mathbf{z}}_{t\mathbf{z},t}-K_{1/t}\right]+\sum_{j<n}\left(\frac{1}{1+\lambda_j t}-\frac{1}{1+\delta t}\right)}{\sum_{j<n}\left(\frac{1}{1+\lambda_j t}-\frac{1}{1+\delta t}\right)}+\mathcal{O}(t-s)\\
&=\frac{\mathrm{Tr}\left[K^{\delta\mathbf{z}}_{t\mathbf{z},t}\right]-\frac{n}{1+\delta t}}{\mathrm{Tr}\left[K_{1/t}\right]-\frac{n}{1+\delta t}}+\mathcal{O}(t-s).
\end{aligned}$$

□

The matrix $K_{1/t}$, that appears in lemma 3.10, has an interpretation in terms of Kirchhoff forests. The entry $K_{1/t}(x,y)$ equals the probability of the event that vertex y is a root of $\Phi_{1/t}$ and that vertex x belongs to the same component as y [19]. In particular, the diagonal entry $K_{1/t}(x,x)$ gives the probability that x is a root. The entries of $K_{1/t}$ also have an interpretation in terms of the killed random walk

$$K_{1/t}(x,y)=\mathbf{P}_x(X_{T_t}=y). \tag{3.49}$$

The matrix $K^{\delta\mathbf{z}}_{t\mathbf{z},t}$ can be interpretated using the Kirchhoff forest $\Phi^\star_{1/t}$ on the extended graph $\mathcal{G}^\star$, which was defined in eq. (3.37), as it holds that

$$K^{\delta\mathbf{z}}_{t\mathbf{z},t}(x,y)=\mathbb{P}(x\leftrightarrow_{\Phi^\star_{1/t}}y,\ y\in\rho^\star_{1/t})+\mathbb{P}(x\leftrightarrow_{\Phi^\star_{1/t}}y,\ (y,\bigstar)\in\Phi^\star_{1/t}). \tag{3.50}$$

That is, $K^{\delta\mathbf{z}}_{t\mathbf{z},t}(x,y)$ denotes the probability of the event that x and y belong to the same component of $\Phi^\star_{1/t}$ and that either y is a root or a neighbor of $\bigstar$.

Lemma 3.11 (PGF of occupation field of closed random walk). *Fix $t>0$. Let Λ_t be an $\mathbb{N}_0^{\mathcal{X}}$-valued random variable with law*

$$P(\Lambda_t=\mathbf{m}):=\mathbf{P}_{\mathrm{unif}}(\ell[(X_k)_{0\le k\le T_t}]=\mathbf{m}\mid X_{T_t+1}=X_0),\qquad \textit{for all } \mathbf{m}\in\mathbb{N}_0^{\mathcal{X}}. \tag{3.51}$$

Then for $\mathbf{z}\in(0,1)^{\mathcal{X}}$ the pgf of Λ_t is given by

$$E\left[\prod_{x\in\mathcal{X}}\mathbf{z}(x)^{\Lambda_t(x)}\right]=\frac{\mathrm{Tr}\left[K^{\delta\mathbf{z}}_{t\mathbf{z},t}\right]-\frac{n}{1+\delta t}}{\mathrm{Tr}\left[K_{1/t}\right]-\frac{n}{1+\delta t}}, \tag{3.52}$$

where $K_{1/t}:=\frac{1}{t}(\frac{1}{t}I+L)^{-1}$ and

$$K^{\delta\mathbf{z}}_{t\mathbf{z},t}:=\left(\left(\tfrac{1}{t}+\delta\right)\mathrm{diag}\left(\tfrac{1-\mathbf{z}}{\mathbf{z}}\right)+\tfrac{1}{t}I+L\right)^{-1}\left(\tfrac{1}{t}\mathrm{diag}\left(\tfrac{1-\mathbf{z}}{\mathbf{z}}\right)+\tfrac{1}{t}I\right). \tag{3.53}$$

Proof. First note that by eq. (3.49), we have that

$$\begin{aligned}\operatorname{Tr}\left[K_{1/t}\right] - \tfrac{n}{1+\delta t} &= \sum_{x\in\mathcal{X}}\left(K_{1/t}(x,x) - \tfrac{1}{1+\delta t}\right) = \sum_{x\in\mathcal{X}}\left(\mathbf{P}_x(X_{T_t}=X_0) - \mathbf{P}_{\mathrm{unif}}(T_t=0)\right)\\ &= \sum_{x\in\mathcal{X}}\mathbf{P}_x(X_{T_t}=X_0,\ T_t\geq 1) = n\mathbf{P}_{\mathrm{unif}}(X_{T_t+1}=X_0)\mathbf{P}_{\mathrm{unif}}(T_t\geq 1).\end{aligned} \tag{3.54}$$

Below we assume that the killed random walk $X=(X_k)_{0\leq k\leq T_t}$ is constructed from the DF-stacks $\{(A_i(x),B_i(x))_{i\in\mathbb{N}_0}\colon\ x\in\mathcal{X}\}$, as described in section 3.1.2.3, with X_0 uniformly distributed on $\mathcal{X}$ and independent of the DF-stacks.

Consider the DF-stacks $\{(A_i^\star(x),B_i^\star(x))_{i\in\mathbb{N}_0}\colon\ x\in\mathcal{X}^\star\}$ on the extended graph $\mathcal{G}^\star$, defined in the proof of lemma 3.9, and recall that they are coupled to the DF-stacks $\{(A_i(x),B_i(x))_{i\in\mathbb{N}_0}\colon\ x\in\mathcal{X}\}$.

For each $x\in\mathcal{X}$ we define three geometric random variables

$$\begin{aligned}T^\bigstar(x) &:= \min\{i\in\mathbb{N}_0\colon\ A_i^\star(x)=\bigstar\},\\ T^\blacklozenge(x) &:= \min\{i\in\mathbb{N}_0\colon\ A_i^\star(x)=\blacklozenge\},\\ T^\dagger(x) &:= \min\{i\in\mathbb{N}_0\colon\ B_i^\star(x)=1\},\end{aligned}$$

and we note that $T^\bigstar(x)\wedge T^\blacklozenge(x)\sim\mathrm{Geom}_{\mathbb{N}_0}(1-\mathbf{z}(x))$.

Using the observation in eq. (3.39) gives us for the pgf of Λ_t that

$$\begin{aligned}E\left[\prod_{x\in\mathcal{X}}\mathbf{z}(x)^{\Lambda_t(x)}\right] &= \mathbb{P}\left(\bigcap_{x\in\mathcal{X}}\{\ell[(X_k)_{0\leq k\leq T_t}](x)\leq T^\bigstar(x)\wedge T^\blacklozenge(x)\}\ \middle|\ X_{T_t+1}=X_0\right)\\ &= \frac{\mathbb{P}\left(\bigcap_{x\in\mathcal{X}}\{\ell[(X_k)_{0\leq k\leq T_t}](x)\leq T^\bigstar(x)\wedge T^\blacklozenge(x)\},\ X_{T_t+1}=X_0\right)}{\mathbb{P}(X_{T_t+1}=X_0)}\\ &= \frac{n\mathbb{P}(T_t\geq 1)\mathbb{P}\left(\bigcap_{x\in\mathcal{X}}\{\ell[(X_k)_{0\leq k\leq T_t}](x)\leq T^\bigstar(x)\wedge T^\blacklozenge(x)\},\ X_{T_t+1}=X_0\right)}{\operatorname{Tr}\left[K_{1/t}\right]-\frac{n}{1+\delta t}}.\end{aligned}$$

Since the denominator is as required, it remains to consider the numerator

$$\begin{aligned}&n\mathbb{P}\left(\bigcap_{x\in\mathcal{X}}\{\ell[(X_k)_{0\leq k\leq T_t}](x)\leq T^\bigstar(x)\wedge T^\blacklozenge(x)\},\ X_{T_t+1}=X_0\right)\mathbb{P}(T_t\geq 1)\\ &= n\mathbb{P}\left(\bigcap_{x\in\mathcal{X}}\{\ell[(X_k)_{0\leq k\leq T_t-1}](x)\leq T^\bigstar(x)\wedge T^\blacklozenge(x)\},\ X_{T_t}=X_0\ \middle|\ T_t\geq 1\right)\mathbb{P}(T_t\geq 1)\\ &= n\mathbb{P}\left(\bigcap_{x\in\mathcal{X}}\{\ell[(X_k)_{0\leq k\leq T_t-1}](x)\leq T^\bigstar(x)\wedge T^\blacklozenge(x)\},\ X_{T_t}=X_0,\ T_t\geq 1\right)\\ &= n\mathbb{P}\left(\bigcap_{x\in\mathcal{X}}\{\ell[(X_k)_{0\leq k\leq T_t}](x)-\mathbf{1}_{\{X_0\}}(x)\leq T^\bigstar(x)\wedge T^\blacklozenge(x)\},\ X_{T_t}=X_0,\ T_t\geq 1\right),\end{aligned}$$

where $\mathbf{1}_{\{X_0\}}\in\{0,1\}^{\mathcal{X}}$ denotes the indicator of X_0.

On the event $\{X_0 = y\}$ the random variable $\mathbf{1}_{\{X_0\}}$ is deterministic, so that $\ell[(X_k)_{0\le k\le T_t-1}]$ and $\mathbf{1}_{\{X_0\}}$ are conditionally independent. Since the pgf of two independent random variables equals the product of their respective pgfs, this gives us that

$$
\begin{aligned}
&n\mathbb{P}\left(\bigcap_{x\in\mathcal{X}}\left\{\ell[(X_k)_{0\le k\le T_t}](x)-\mathbf{1}_{\{X_0\}}(x)\le T^\bigstar(x)\wedge T^\blacklozenge(x)\right\},\ X_{T_t}=X_0,\ T_t\ge 1\right)\\
&=\sum_{y\in\mathcal{X}}\mathbb{P}\left(\bigcap_{x\in\mathcal{X}}\left\{\ell[(X_k)_{0\le k\le T_t}](x)-\mathbf{1}_{\{y\}}(x)\le T^\bigstar(x)\wedge T^\blacklozenge(x)\right\},\ X_{T_t}=y,\ T_t\ge 1\ \middle|\ X_0=y\right)\\
&=\sum_{y\in\mathcal{X}}\frac{\mathbb{P}\left(\bigcap_{x\in\mathcal{X}}\left\{\ell[(X_k)_{0\le k\le T_t}](x)\le T^\bigstar(x)\wedge T^\blacklozenge(x)\right\},\ X_{T_t}=y,\ T_t\ge 1\ \middle|\ X_0=y\right)}{\mathbb{P}\left(1\le T^\bigstar(y)\wedge T^\blacklozenge(y)\right)}\\
&=\sum_{y\in\mathcal{X}}\mathbf{z}(y)\mathbb{P}\left(\bigcap_{x\in\mathcal{X}}\left\{\ell[(X_k)_{0\le k\le T_t}](x)\le T^\bigstar(x)\wedge T^\blacklozenge(x)\right\},\ X_{T_t}=y,\ T_t\ge 1\ \middle|\ X_0=y\right).
\end{aligned}
$$

Defining

$$
\begin{aligned}
\tau^\bigstar &:= \min\{m\in\mathbb{N}_0\colon\ \exists x\in\mathcal{X}\text{ s.t. }\ell[(X_k)_{0\le k\le m}](x)>T^\bigstar(x)\},\\
\tau^\blacklozenge &:= \min\{m\in\mathbb{N}_0\colon\ \exists x\in\mathcal{X}\text{ s.t. }\ell[(X_k)_{0\le k\le m}](x)>T^\blacklozenge(x)\},
\end{aligned}
$$

gives us that

$$
\bigcap_{x\in\mathcal{X}}\{\ell[(X_k)_{0\le k\le T_t}](x)\le T^\bigstar(x)\wedge T^\blacklozenge(x)\}=\{T_t<\tau^\bigstar\wedge\tau^\blacklozenge\}. \tag{3.55}
$$

On the event $\{T_t<\tau^\bigstar\wedge\tau^\blacklozenge\}$ we have that

$$
T_t=\min\{m\in\mathbb{N}_0\colon\ \exists x\in\mathcal{X}\text{ s.t. }\ell[(X_k)_{0\le k\le m}](x)>T^\dagger(x)\}, \tag{3.56}
$$

from which follows that

$$
\begin{aligned}
&\sum_{y\in\mathcal{X}}\mathbf{z}(y)\mathbb{P}\left(\bigcap_{x\in\mathcal{X}}\{\ell[(X_k)_{0\le k\le T_t}](x)\le T^\bigstar(x)\wedge T^\blacklozenge(x)\},\ X_{T_t}=y,\ T_t\ge 1\ \middle|\ X_0=y\right)\\
&\qquad=\sum_{y\in\mathcal{X}}\mathbf{z}(y)\mathbb{P}\left(1\le T_t<\tau^\bigstar\wedge\tau^\blacklozenge,\ X_{T_t}=y\mid X_0=y\right)\\
&\qquad=\sum_{y\in\mathcal{X}}\mathbf{z}(y)\mathbb{P}\left(1\le T_t<\tau^\bigstar\wedge\tau^\blacklozenge,\ X_{T_t}=y,\ T^\dagger(y)<T^\bigstar(y)\ \middle|\ X_0=y\right)\\
&\qquad=\sum_{y\in\mathcal{X}}\mathbf{z}(y)\mathbb{P}\left(1\le T_t\wedge\tau^\bigstar<\tau^\blacklozenge,\ X_{T_t\wedge\tau^\bigstar}=y,\ T^\dagger(y)<T^\bigstar(y)\ \middle|\ X_0=y\right).
\end{aligned}
$$

By independence we find that

$$\begin{aligned}
&\sum_{y\in\mathcal{X}} \mathbf{z}(y)\mathbb{P}\left(1 \le T_t \wedge \tau^\star < \tau^\blacklozenge,\ X_{T_t\wedge\tau^\star} = y,\ T^\dagger(y) < T^\star(y) \;\middle|\; X_0 = y\right) \\
&= \sum_{y\in\mathcal{X}} \mathbf{z}(y)\mathbb{P}\left(1 \le T_t \wedge \tau^\star < \tau^\blacklozenge,\ X_{T_t\wedge\tau^\star} = y \mid X_0 = y\right)\mathbb{P}\left(T^\dagger(y) < T^\star(y)\right) \\
&= \sum_{y\in\mathcal{X}} \mathbb{P}\left(1 \le T_t \wedge \tau^\star < \tau^\blacklozenge,\ X_{T_t\wedge\tau^\star} = y \mid X_0 = y\right) \\
&= \sum_{y\in\mathcal{X}} \left(\mathbb{P}\left(T_t \wedge \tau^\star < \tau^\blacklozenge,\ X_{T_t\wedge\tau^\star} = y \mid X_0 = y\right) - \tfrac{1}{1+\delta t}\right).
\end{aligned}$$

Using eq. (3.50) gives us that

$$\begin{aligned}
&\sum_{y\in\mathcal{X}} \left(\mathbb{P}\left(T_t \wedge \tau^\star < \tau^\blacklozenge,\ X_{T_t\wedge\tau^\star} = y \mid X_0 = y\right) - \tfrac{1}{1+\delta t}\right) \\
&= \sum_{y\in\mathcal{X}} \left(\mathbb{P}\left((y,\star) \in \Phi^\star_{1/t} \text{ or } y \in \rho^\star_{1/t}\right) - \tfrac{1}{1+\delta t}\right) \\
&= \sum_{y\in\mathcal{X}} \left(K^{\delta\mathbf{z}}_{t\mathbf{z},t}(y,y) - \tfrac{1}{1+\delta t}\right) = \operatorname{Tr}\left[K^{\delta\mathbf{z}}_{t\mathbf{z},t}\right] - \tfrac{n}{1+\delta t},
\end{aligned}$$

which completes the proof. □

Proof of theorem 3.6. The result of theorem 3.6 follows from lemmas 3.7, 3.8, 3.10 and 3.11. □

3.4.4 Closed walk decomposition

Corollary 3.6.1 is a simple consequence of theorem 3.6.

Proof of corollary 3.6.1. Fix $\gamma = (x_0, x_1, \dots, x_l) \in \mathcal{P}^{\mathrm{cl}}$. By theorem 3.6 the intensity measure μ_γ is given by

$$\mu_\gamma((a,b]) = \int_a^b \kappa_t \mathbf{P}_{\mathrm{unif}}((X_k)_{0\le k\le T_t+1} = \gamma \mid X_{T_t+1} = X_0)\ \mathrm{d}t.$$

Write

$$\mathbf{P}_{\mathrm{unif}}((X_k)_{0\le k\le T_t+1} = \gamma \mid X_{T_t+1} = X_0) = \frac{\mathbf{P}_{\mathrm{unif}}((X_k)_{0\le k\le T_t+1} = \gamma)}{\mathbf{P}_{\mathrm{unif}}(X_{T_t+1} = X_0)}.$$

For the denominator we have by eq. (3.54) that

$$\mathbf{P}_{\mathrm{unif}}(X_{T_t+1} = X_0) = \frac{1+\delta t}{n\delta t}\left(\operatorname{Tr}\left[K_{1/t}\right] - \frac{n}{1+\delta t}\right) = \frac{(1+\delta t)\kappa_t}{n\delta}. \tag{3.57}$$

We therefore have that

$$\begin{aligned}
&\mathbf{P}_{\mathrm{unif}}((X_k)_{0\le k\le T_t+1} = \gamma \mid X_{T_t+1} = X_0) = \frac{n\delta}{(1+\delta t)\kappa_t}\mathbf{P}_{\mathrm{unif}}((X_k)_{0\le k\le T_t+1} = \gamma) \\
&= \frac{\delta}{(1+\delta t)^2\kappa_t}\left(\frac{\delta t}{1+\delta t}\right)^{l-1}\mathbf{P}_{x_0}((X_k)_{0\le k\le l} = \gamma) \\
&= \frac{\delta}{(1+\delta t)^2\kappa_t}\left(\frac{\delta t}{1+\delta t}\right)^{l-1}\prod_{i=1}^{l}\frac{w(x_{i-1},x_i)}{\delta} = \frac{1}{(1+\delta t)^2\kappa_t}\left(\frac{t}{1+\delta t}\right)^{l-1} w(e(\gamma)),
\end{aligned}$$

which completes the proof. □

CHAPTER 4

Coupled Kirchhoff forests: dynamic loop-ensemble, Laplacian spectrum & DGFF

This chapter is based on joint work in progress with L. Avena and A. Gaudillière.

4.1 Introduction

This chapter is a continuation of chapter 3. As in chapter 3 we consider the *coupled forest process*, defined in section 3.2, that is constructed by coupling together a continuum of realizations of Wilson's algorithm. Where in chapter 3 the observable of interest was the associated occupation field, in the current chapter the focus is on the configurations of colored cycles that are 'popped' (i.e. deleted) during Wilson's procedure. In this chapter, three results will be presented that are related to these cycle configurations.

(1) *Poissonian loop-ensemble.* In [55] Le Jan introduced a Poisson point process on the set of closed walk, called the Poissonian loop-ensemble, and showed how such a loop-ensemble can be obtained from a single application of Wilson's algorithm by using Poisson-Dirichlet random variables to decompose the closed walks generated by the algorithm.

The first result of this chapter, theorem 4.1, shows that using the dynamic coupled version of Wilson's algorithm circumvents the need for Poisson-Dirichlet decompositions, allowing for the construction of the loop-ensemble directly from the cycle popping procedure, without requiring any additional randomness.

An important role in this construction is being filled by a bijection between the set of closed walks and a subset of 'rooted' configurations of colored cycles. This bijection, which is defined in section 4.5.1 below, allows us to associate a closed walk to each time of the coupled forest process at which new cycles are popped. Incidentally, in proving theorem 4.1 we will provide an alternative proof of theorem 3.6, which is more probabilistic in flavor than the generating function based proof provided in the previous chapter.

(2) *Spectral decomposition.* The second result, provided in theorem 4.2 and its direct corollary corollary 4.2.1, shows that under some assumptions on the Laplacian, the occupation field process admits a spectral decomposition, in the sense that its distribution can be written as a mixture of which each term corresponds to a single Laplacian eigenvalue.

Recently, in [9], the coupled forest process was used to devise an estimation procedure for the Laplacian spectrum of a graph. The procedure devised there utilizes an observable obtained solely from the forests, that is distributed as a sum of random variables, that are each associated with one of the Laplacian eigenvalues. In contrast, the spectral decomposition provided in theorem 4.2 shows that, by considering the cycle configurations produced by Wilson's algorithm rather than the forests, we can construct observables that have a mixture distribution, with each component of the mixture depending on a single eigenvalue. As mixtures are better suited than sums for estimating the parameters of the components, theorem 4.2 could possibly be used to improve on the estimation procedure in [9].

(3) *Gaussian free field with mass.* Our third result concerns the relation between the occupation field of Wilson's algorithm and the Gaussian free field. Lupu exhibited in [59] a remarkable coupling between the discrete Gaussian free field and the Poissonian loop-ensemble. This coupling fits very well with the dynamic framework of the coupled

forest process. Adjusting Lupu's coupling to this framework, allows us to couple together in a continuum of Gaussian free fields, parametrized by their distinct masses, as is shown in proposition 4.3.

4.1.1 Colored cycle configurations

In this section we start by providing a definition of the random colored cycle configurations produced by Wilson's algorithm. As this chapter is a continuation of chapter 3, the setting in the this chapter will be identical to the setting presented in section 3.1.1, which therefore will not be repeated here.

Recall from section 3.1.2.4 that for fixed $t > 0$ cycle popping of the DF-stacks $\{(A_i(x), B_i^t(x))_{i\in\mathbb{N}_0} : \ x \in \mathcal{X}\}$ produces a sequence of 'popped' cycles $(\Gamma_0^\circ, \Gamma_1^\circ, \dots, \Gamma_{i^*-1}^\circ)$, where, conditionally on the DF-stacks, the exact sequence that is produced depends on some arbitrary choice of ordering on the set of cycles. This dependence of the sequence of cycles on an arbitrary ordering is inconvenient. Therefore, we will instead consider the set of colored cycles produced by cycle popping, since, as was shown by Wilson [77, Thm. 4], the set of colored cycles produced is the same for any choice of ordering of the cycles used in the cycle popping procedure.

A *colored cycle*[1] is a pair (γ°, c), where γ° is a cycle and $c : s(\gamma^\circ) \to \mathbb{N}_0$ is a map that assigns an integer 'color' to each vertex in the cycle.

By giving each vertex in a popped cycle Γ_i° a color value equal the number of deleted arrows from its stack, we see that the cycle popping procedure produces a set of colored cycles $\{(\Gamma_i^\circ, c_i) : \ i \in [i^* - 1]_0\}$, with the coloring map $c_i : s(\gamma^\circ) \to \mathbb{N}_0$ given by $c_i(x) := \mathbf{d}_i(x)$, where we recall from section 3.1.2.4 that $\mathbf{d}_i(x)$ denotes the number of deleted arrows. So, an arrow's color in a colored cycle corresponds to its 'level' in the DF-stack.

For $t \geq 0$ we denote by $\mathfrak{C}_t$ the set of colored cycles produced by cycle popping the DF-stacks $\{(A_i(x), B_i^t(x))_{i\in\mathbb{N}_0} : \ x \in \mathcal{X}\}$. We note that $\mathfrak{C}_s \subseteq \mathfrak{C}_t$ for $s < t$, since as time progresses an increasing amount of cycles are popped from the DF-stacks. Recalling from section 3.2.1 that $(N_t)_{t\geq 0}$ denotes the occupation field process, we note that the occupation field N_t can be expressed using the colored cycles $\mathfrak{C}_t$ as

$$N_t = \underline{1} + \sum_{(\gamma^\circ, c)\in\mathfrak{C}_t} \mathbf{1}_{s(\gamma^\circ)}. \tag{4.1}$$

That is, N_t is one more than the sum, taken over all cycles occurring in $\mathfrak{C}_t$, of the indicators of their support.

[1] This terminology was introduced by Wilson and is adopted here. We assume no confusion arises due to the red and green colors of the arrows introduced in section 3.1.2.4.

4.2 Constructing the Poissonian loop-ensemble

The Poissonian loop-ensemble is associated to a given sub-Markovian transition matrix. Here we give the definition of the loop-ensemble for the specific sub-Markovian transition matrix $(I - \frac{1}{1/t+\delta}L)$ of the killed random walk $(X_k)_{0\le k\le T_t}$ introduced in section 3.1.1. We remark that, in principle, the coupled forest process can be constructed using other choices of random walks, and can accommodate inhomogeneous killing rates among vertices as well. Hence, our specific choice of random walk should not be seen as a restriction.

As was done in section 3.1.2.1, we denote by $\mathcal{P}^{\mathrm{cl}}$ the set of closed walks.

Definition 4.2.1. For $\alpha > 0$ a Poisson point process on $\mathcal{P}^{\mathrm{cl}}$ with intensity measure

$$\{\gamma\} \mapsto \frac{\alpha}{l}\frac{w(e(\gamma))}{(1/t+\delta)^l}, \qquad \text{for all } \gamma = (x_0, x_1, \dots, x_l) \in \mathcal{P}^{\mathrm{cl}} \tag{4.2}$$

is called a *loop-ensemble* of intensity α with killing rate $1/t$. ■

This definition differs from that of Le Jan, in that it omits the infinitely many length 0 walks (trivial loops) that occur in the loop-ensemble according to the definition by Le Jan.

The theorem below introduces a dynamic generalization $\mathcal{L}$ of the loop-ensemble, and shows that it can be constructed with the Kirchhoff forest coupling introduced in section 3.2.

Theorem 4.1. *There exists a Poisson point process $\mathcal{L}$ on the product space $\mathcal{P}^{\mathrm{cl}} \times (0,\infty)$ such that for any $t > 0$ the random atomic measure on $\mathcal{P}^{\mathrm{cl}}$ defined by*

$$\{\gamma\} \mapsto \mathcal{L}(\{\gamma\} \times (0,t]) \tag{4.3}$$

is a loop-ensemble of intensity 1 with killing rate $1/t$, that is measurable with respect to the σ-algebra generated by $(\mathfrak{C}_s, \Phi_{1/s})_{0\le s\le t}$.

The measurabiltity statement in theorem 4.1 can be made explicit. The process $\mathcal{L}$ is obtained from $(\mathfrak{C}_s, \Phi_{1/s})_{0\le s\le t}$ by applying a specific bijection, defined in lemma 4.4, between cycle configurations and closed walks, as will be detailed in the proof of theorem 4.1 below.

4.3 Spectral decomposition

Recall from eq. (3.28) the definition of the set of jump times

$$\mathcal{T} := \{t \in (0,\infty) \colon \quad N_t \neq \lim_{s\uparrow t} N_s\}. \tag{4.4}$$

The jump rate κ_t in eq. (3.27) can be written as a sum with each term depending only on a single eigenvalue of L,

$$\kappa_t := \mathrm{Tr}\left[\tfrac{1}{t}(K_{1/t} - \tfrac{1}{1+\delta t}I)\right] = \sum_{j<n} \frac{\delta - \lambda_j}{(1+\delta t)(1+\lambda_j t)}. \tag{4.5}$$

This formulation of κ_t suggest that $\mathcal{T}$ can be interpreted as a union of independent Poisson point processes where each Poisson point process is associated to an eigenvalue.

We require two additional assumptions on the Laplacian matrix L and the random walk parameter δ.

ASSUMPTION 1a:

The random walk X is irreducible and reversible. (AS1a)

From the irreducibility assumption follows that X has a unique stationary distribution, denoted by μ. The assumption (AS1a) on the Laplacian further ensures that the Laplacian spectrum is real valued, and non-negative. The maximal Laplacian eigenvalue is denoted by $\lambda_{\max}$. Moreover, with this assumption there exists a basis $\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{n-1}$ of $\mathbb{R}^{\mathcal{X}}$ consisting of eigenvectors of L that are orthonormal with respect to the inner product $\langle f, g\rangle_\mu := \sum_{x\in\mathcal{X}} f(x)g(x)\mu(x)$. Naturally, these eigenvectors are indexed in such a manner that $\mathbf{v}_j$ has corresponding eigenvalue λ_j.

Our second assumption adds a restriction to the parameter δ.

ASSUMPTION 2:

The random walk parameter satisfies $\delta \geq \lambda_{\max}$. (AS2)

By Gershgorin's circle theorem it holds that $\lambda_{\max} \leq \max_{x\in\mathcal{X}} 2L(x,x)$, so that a sufficient condition for assumption (AS2) to hold is that $\delta \geq \max_{x\in\mathcal{X}} 2L(x,x)$.

The two assumptions (AS1a) and (AS2) are chosen to ensure that for all $j < n$ the individual terms $\frac{\delta-\lambda_j}{(1+\delta t)(1+\lambda_j t)}$ of the Poisson rate function in eq. (4.5) are real valued and non-negative. This allows us to define for all $j < n$ and $t \in [0,\infty)$ the random geometric killing times T_j^t, the random vertices Y_j, and for all $x \in \mathcal{X}$, $l \in \mathbb{N}$ the random closed walks $\Gamma_{l,x}$, such that all these random variables are independent with joint law P given by

$$\begin{aligned} P(\Gamma_{l,x} = \gamma) &:= \mathbf{P}_x((X_k)_{0\leq k\leq l} = \gamma \mid X_l = x) = \frac{w(e(\gamma))}{\delta^l(I - \frac{1}{\delta}L)^l(x,x)}, \\ P(T_j^t = k) &:= \frac{1+\lambda_j t}{1+\delta t}\left(1 - \frac{1+\lambda_j t}{1+\delta t}\right)^k, \qquad P(Y_j = x) := \mathbf{v}_j(x)^2\mu(x), \end{aligned} \tag{4.6}$$

where l denotes the length of γ.

Theorem 4.2. *The Poisson point process $\mathcal{L}$ on $\mathcal{P}^{\mathrm{cl}} \times (0,\infty)$ has intensity measure*

$$\{\gamma\} \times (a,b] \mapsto \sum_{j<n} \int_a^b \tfrac{\delta-\lambda_j}{(1+\delta t)(1+\lambda_j t)}\; P(Y_j = x_0)P(T_j^t + 1 = l)P(\Gamma_{l,x_0} = \gamma)\; \mathrm{d}t,$$

with $\gamma = (x_0, x_1, \ldots, x_l) \in \mathcal{P}^{\mathrm{cl}}$.

So, at each jump time the closed walk that is generated by the cycle popping procedure by applying the bijection of lemma 4.4, has the same law as a closed walk obtained as follows. First sample a random eigenvalue with probability proportional to $\frac{\delta-\lambda_j}{(1+\delta t)(1+\lambda_j t)}$, then conditionally on the eigenvalue sample independently a random starting point and a random length, and finally sample a closed walk with the given starting point and length.

A remarkable aspect of theorem 4.2 is that, although the increment of the running time $\sum_{x\in\mathcal{X}}\ell[\mathfrak{C}_t^{\Delta}](x)$ and the waking root V_t are themselves not independent of each other, by conditioning on the 'artificial' random eigenvalue their counterparts T_j^t and Y_j do become independent.

As a consequence of theorem 4.2, we obtain, in corollary 4.2.1 below, a rephrasing of the result of the previous chapter.

On the same probability space where we defined the random variables Y_j and T_j^t of eq. (4.6), we further define for all $j < n$ Poisson point processes Ψ_j with intensity measures

$$\mu_j((a,b]) := \int_a^b \frac{\delta-\lambda_j}{(1+\delta t)(1+\lambda_j t)}\,\mathrm{d}t, \tag{4.7}$$

and for all $t > 0$ we let $\Gamma_{l,x,t}$ be copies of $\Gamma_{l,x}$, such that all these random variables are independent.

Corollary 4.2.1. *The distributions of the running time process and of the occupation field process are, respectively, given by*

$$(M_t)_{t\geq 0} \overset{d}{=} \left(1+\sum_{j<n}\int_0^t (T_j^s+1)\ \Psi_j(\mathrm{d}s)\right)_{t\geq 0} \tag{4.8}$$

and

$$(N_t)_{t\geq 0} \overset{d}{=} \left(\underline{1}+\sum_{j<n}\int_0^t \ell[\Gamma^-_{T_j^s+1,Y_j,s}]\ \Psi_j(\mathrm{d}s)\right)_{t\geq 0}. \tag{4.9}$$

4.4 Coupling DGFFs of different masses

In this section we can drop the assumption (AS2) above. Instead, we require a strengthening of assumption (AS1a).

ASSUMPTION 1b:

The Laplacian matrix L is symmetric. (AS1b)

Assumption (AS1b) ensures that the Gaussian free field, defined in definition 4.4.1 below, is well-defined.

Definition 4.4.1 (Discrete Gaussian free field with mass). For $q > 0$, an n-dimensional centered Gaussian random variable ϕ with covariance matrix

$$\Sigma_q^2 := \tfrac{1}{2}(q+\delta)(qI+L)^{-1} \tag{4.10}$$

is called a *discrete Gaussian free field* (abbrv. DGFF) on $\mathcal{G}$ with mass q. ■

The above definition is unconventional, as customarily the scaling factor $\frac{1}{2}(q+\delta)$ is omitted from the defining covariance matrix. However, the use of this scaling factor is better suited to our purposes.

As was shown by Le Jan in [56], the occupation field of Wilson's algorithm is related to the Gaussian free field. To explore this connection, we add exponential waiting times to the random walks used in Wilson's algorithm.

To the DF-stacks used to define the coupled forest process we add stacks of exponential waiting times, thus obtaining the collection $\{(A_i(x), U_i(x), \eta_i(x))_{i\in\mathbb{N}_0} :\ x \in \mathcal{X}\}$ where for each $k \in \mathbb{N}_0$, $x \in \mathcal{X}$ the random variables $\eta_k(x)$ are Exp(1) distributed and independent of each other and of all $A_i(x)$ and $U_i(x)$. We define the *continuous occupation field process* $(\Xi_t)_{t\geq 0}$ as the $(0,\infty)^{\mathcal{X}}$-valued process

$$\Xi_t(x) := \sum_{k=0}^{N_t(x)-1} \eta_k(x). \tag{4.11}$$

The scaled continuous occupation field $\frac{t}{1+\delta t}\Xi_t$ corresponds to the occupation field of Wilson's algorithm at intensity $1/t$, when we use continuous-time random walks with infinitesimal generator $-L$, killed at random exponential times of rate $1/t$. While this scaled field might be a more natural object than Ξ_t, the process $(\Xi_t)_{t\geq 0}$ has the convenient property of having piece-wise constant trajectories.

From definition (4.11) it follows that $\Xi_0(x) = \eta_0(x)$. By the Box-Muller transform, we can assume for every $x \in \mathcal{X}$ that $\eta_0(x)$ is coupled to two independent standard Gaussians $Z(x)$ and $\tilde{Z}(x)$ such that $\eta_0(x) = \frac{1}{2}Z^2(x) + \frac{1}{2}\tilde{Z}^2(x)$, where we write $Z^2 := (Z(x)^2)_{x\in\mathcal{X}}$. Thus allowing us to decompose Ξ_0 as a sum of squares two i.i.d Gaussians.

For each $t > 0$ let $Y_t \sim \mathrm{Ber}(\frac{1}{2})$ be Bernoulli random variables, that are independent of each other and of the DF-stacks $\{(A_i(x), U_i(x), \eta_i(x))_{i\in\mathbb{N}_0} :\ x \in \mathcal{X}\}$, and denote by $\Xi_t^\Delta := \Xi_t - \lim_{s\uparrow t}\Xi_s$ the increment of the continuous occupation field at time t.

Denote by $\Psi := \sum_{\tau\in\mathcal{T}}\delta_\tau$ the random atomic measure supported on $\mathcal{T}$, the set of jump times of the occupation field process.

We then define a *thinned occupation field process* $(\xi_t)_{t\geq 0}$ by

$$\xi_t := \sqrt{\tfrac{1}{2}Z^2 + \int_0^t \Xi_s^\Delta Y_s \ \Psi(\mathrm{d}s)}. \tag{4.12}$$

We remark that this construction provides us with a second process $(\tilde{\xi}_t)_{t\geq 0}$, that is defined by

$$\tilde{\xi}_t := \sqrt{\tfrac{1}{2}\tilde{Z}^2 + \int_0^t \Xi_s^\Delta (1-Y_s) \ \Psi(\mathrm{d}s)}. \tag{4.13}$$

Since the processes $(\xi_t)_{t\geq 0}$ and $(\tilde{\xi}_t)_{t\geq 0}$ are obtained by thinning the Poisson point process Ψ, they are independent of each other.

Le Jan showed [56, Thm. 2] that the thinned occupation field is distributed as the absolute value of a DGFF with mass $1/t$, i.e.

$$\xi_t \overset{d}{=} |\phi|_t, \tag{4.14}$$

where ϕ_t is a DGFF with mass $1/t$, and $|\phi|_t$ denotes its term-wise absolute value, $|\phi|_t(x) := |\phi_t(x)|$ for all $x \in \mathcal{X}$.

In addition to the thinned occupation field $(\xi_t)_{t\geq 0}$, we further require for each *undirected* edge $e = \{x, y\}$ independent uniform random variables $U_e \sim \mathrm{Unif}(0, 1)$. We recall that L is symmetric, so that the set of undirected edges is given by $\bar{\mathcal{E}} := \{\{x, y\} \in \binom{\mathcal{X}}{2} \colon\ (x, y) \in \mathcal{E}\}$. The variable U_e is used to couple together for all $t > 0$ the Bernoulli random variables O_e^t defined by

$$O_e^t := \mathbf{1}\left\{U_e > \exp\left(-\tfrac{2tw(x,y)}{1+\delta t}\sqrt{\xi_t(x)\xi_t(y)}\right)\right\}.$$

These Bernoulli variables were introduced in Lupu's construction of the DGFF [59, Thm. 1], and are adapted here to the dynamic setting by coupling them together for distinct t.

Recall that $\xi_0 := \frac{1}{\sqrt{2}}|Z|$, where Z is a vector of i.i.d. standard Gaussians, and write $S_0 := \mathrm{sign}(Z)$, i.e. $S_0(x) := \mathrm{sign}(Z(x))$ for all $x \in \mathcal{X}$.

Define the filtration $(\mathcal{F}_t)_{t\geq 0}$ by $\mathcal{F}_t := \sigma(\xi_s, (O_e^s)_{e\in\bar{\mathcal{E}}}, S_0 \colon\ s \leq t)$.

Proposition 4.3. *There exists an $\mathbb{R}^{\mathcal{X}}$-valued stochastic process $(\phi_t)_{t\geq 0}$ with the following properties:*

- *(i) for each $t > 0$ the marginal ϕ_t is a DGFF with mass $1/t$;*
- *(ii) trajectories of $(\phi_t)_{t\geq 0}$ are piece-wise constant and càdlàg;*
- *(iii) the process $(\phi_t)_{t\geq 0}$ satisfies the Markov property;*
- *(iv) the same sign components, i.e. connected components of the spanning subgraph with edges $\{(x, y) \in \mathcal{E} \colon\ \mathrm{sign}(\phi_t(x)) = \mathrm{sign}(\phi_t(y))\}$, are unions of coalescing clusters;*
- *(v) for each $x \in \mathcal{X}$ the map $t \mapsto |\phi_t(x)|$ is non-decreasing;*
- *(vi) the process $(\phi_t)_{t\geq 0}$ is adapted to the filtration $(\mathcal{F}_t)_{t\geq 0}$.*

Proof. For each $t \geq 0$, we will set the absolute value of ϕ_t equal to ξ_t, hence we only require a procedure to determine $\mathrm{sign}(\phi_t)$.

For each $t \geq 0$ we define a spanning subgraph $\mathcal{G}_t = (\mathcal{X}, \mathcal{E}_t)$, whose edges are given by

$$\mathcal{E}_t := \bigcup_{(\gamma^\circ,\, c)\in\mathfrak{C}_t} e(\gamma^\circ) \cup \{(x, y) \in \mathcal{E} \colon\ O^t_{\{x,y\}} = 1\},$$

and denote by $\mathcal{W}_t$ the set of connected components of $\mathcal{G}_t$. Note that the process $(\mathcal{W}_t)_{t\geq 0}$ is a process of coalescing partitions, such that at time $t = 0$ the starting

partition $\mathcal{W}_0$ consists of isolated vertices.

We let $\mathcal{S} := \{t \in (0,\infty)\colon\ \mathcal{W}_t \neq \lim_{s\uparrow t} \mathcal{W}_s\}$ be the set of jump times of the process $(\mathcal{W}_t)_{t\geq 0}$. Write $\tau_0 := 0$ and $\mathcal{S} = \{\tau_1, \tau_2, \tau_3, \ldots\}$ with $\tau_i < \tau_{i+1}$ for all $i \in \mathbb{N}_0$.

For each $\tau_i \in \mathcal{S}$ we will construct a $\{-1,1\}^{\mathcal{X}}$-valued sign vector S_{τ_i} at time τ_i, that assigns the same sign to vertices belonging to the same block of the partition $\mathcal{W}_{\tau_i}$. For this we require some arbitrary decision procedure that can be used to iteratively determine S_{τ_i}, given the partition $\mathcal{W}_{\tau_i}$, the previous signs $S_{\tau_{i-1}}$, and the absolute values $\xi_{\tau_{i-1}}$. An example of such a procedure would be, to assign to a block in $\mathcal{W}_{\tau_i}$ that is obtained by the coalescing of several blocks of $\mathcal{W}_{\tau_{i-1}}$, the sign of the largest of these coalescing blocks.

Denoting by $\Pi(\mathcal{X})$ the set of partitions of $\mathcal{X}$, we can represent this decision procedure by a map $g : \Pi(\mathcal{X}) \times (0,\infty)^{\mathcal{X}} \times \{-1,1\}^{\mathcal{X}} \times (0,\infty) \to \{-1,1\}^{\mathcal{X}}$ such that $g(\pi, \mathbf{x}, \sigma, t)(x) = g(\pi, \mathbf{x}, \sigma, t)(y)$ for any x, y that belong to the same block of the partition π. We then set

$$S_{\tau_i} := g(\mathcal{W}_{\tau_i}, \xi_{\tau_{i-1}}, S_{\tau_{i-1}}, \tau_i),$$

where we recall that $S_{\tau_0} = \operatorname{sign}(Z)$.

Having constructed the signs S_{τ_i} at all times in $\mathcal{S}$, we construct S_t for any $t > 0$ by setting

$$S_t := S_{\tau_{i_t}}, \qquad \text{where } i_t := \max\{i \in \mathbb{N}_0\colon\ \tau_{i_t} \leq t\}.$$

We then define the process $(\phi_t)_{t\geq 0}$ by

$$\phi_t(x) := S_t(x)\xi_t(x).$$

That ϕ_t is a DGFF with mass $1/t$ follows directly from Lupu's coupling [59, Thm. 1]. Properties $(ii) - (vi)$ follow by construction. □

4.5 Proofs

4.5.1 Bijection between closed walks and popped cycles

We call a finite set of colored cycles *admissible* if it could be produced by cycle popping some realization of the DF-stacks. Equivalently, a set C of colored cycles is admissible if it holds that:

- (i) for all distinct $(\gamma_1^\circ, c_1), (\gamma_2^\circ, c_2) \in C$ either $c_1(x) < c_2(x)$ for all $x \in s(\gamma_1^\circ) \cap s(\gamma_2^\circ)$ or $c_1(x) > c_2(x)$ for all $x \in s(\gamma_1^\circ) \cap s(\gamma_2^\circ)$;
- (ii) for all $(\gamma^\circ, c) \in C$ and any $x \in s(\gamma^\circ)$ with $c(x) \geq 1$, there exists $(\tilde{\gamma}^\circ, \tilde{c}) \in C$ with $x \in s(\tilde{\gamma}^\circ)$ and $\tilde{c}(x) = c(x) - 1$;
- (iii) if for any $k \in \mathbb{N}_{\geq 3}$ and $(\gamma_1^\circ, c_1), \ldots, (\gamma_k^\circ, c_k) \in C$ it holds that $c_i(x) < c_{i+1}(x)$ for all $i \in [k-1]$ and all $x \in s(\gamma_i^\circ) \cap s(\gamma_{i+1}^\circ)$, then it holds that $c_1(x) < c_k(x)$ for all $x \in s(\gamma_1^\circ) \cap s(\gamma_k^\circ)$.

The set of all admissible sets of colored cycles is denoted by $\mathcal{A}$.

Wilson observed that for a fixed multi-set of cycles there is a bijection between its admissible colorings and the partial orderings on that multi-set. From an admissible set of colored cycles C we obtain a partial ordering $\preccurlyeq$ by writing $\tilde{\gamma}^\circ \preccurlyeq \gamma^\circ$ for any $(\gamma^\circ, c), (\tilde{\gamma}^\circ, \tilde{c}) \in C$ if (γ°, c) has to be popped before $(\tilde{\gamma}^\circ, \tilde{c})$ can be popped during the cycle popping procedure. That is, if there exist $k \in \mathbb{N}_0$ and $(\gamma_0^\circ, c_0), \dots, (\gamma_k^\circ, c_k) \in C$ with $(\gamma_0^\circ, c_0) = (\gamma^\circ, c)$ and $(\gamma_k^\circ, c_k) = (\tilde{\gamma}^\circ, \tilde{c})$ such that for all $i \in [k]$ it holds that $s(\gamma_{i-1}^\circ) \cap s(\gamma_i^\circ) \neq \varnothing$ and $c_{i-1}(x) < c_i(x)$ for all $x \in s(\gamma_{i-1}^\circ) \cap s(\gamma_i^\circ)$.

The inverse of the above bijection is as follows. For each cycle γ° in partially ordered multi-set of cycles let its coloring map c be such that $c(x)$ equals the number of cycles with x in their support that are smaller than γ°.

So, cycles that are colored using larger values will be smaller in the resulting partial order. This reflects that cycles with large colors appear lower in the DF-stack.

We call an admissible set of colored cycles C a *cycle clump* if it contains a unique monochromatic cycle with color 0, i.e. if

$$|\{(\gamma^\circ, c) \in C\colon\ c(x) = 0 \text{ for all } x \in s(\gamma^\circ)\}| = 1. \tag{4.15}$$

The cycle with color 0 we denote by $\gamma_{\max}^\circ$. This notation refers to the fact that the partially ordered multi-set of cycles associated to a cycle clump has $\gamma_{\max}^\circ$ as its maximum. A pair (C, v) consisting of a cycle clump C and a vertex $v \in \mathcal{X}$ is called a *rooted cycle clump* if $v \in s(\gamma_{\max}^\circ)$.

Lemma 4.4. *There exists an explicit bijection f from the set of rooted cycle clumps to the set of closed walks $\mathcal{P}^{\mathrm{cl}}$ such that for any rooted cycle clump (C, v) it holds for $f(C, v) = (x_0, x_1, \dots, x_l)$ that*

(i) $x_0 = x_l = v$;

(ii) $\sum_{(\gamma^\circ, c) \in C} \mathbf{1}_{e(\gamma^\circ)} = \sum_{i \in [l]} \mathbf{1}_{\{(x_{i-1}, x_i)\}}$.

That is, v is the starting point of the closed walk $f(C, v)$, and the edges contained in the cycle clump C are the same as the edges traversed by the closed walk.

Proof. We use the above defined partial ordering to define the inverse $\overleftrightarrow{C}$ of an admissible set of colored cycles C. If $\preccurlyeq$ is the partial ordering associated with C, then we let $\overleftrightarrow{C}$ be the admissible set of colored cycles associated with the reversed partial ordering $\overleftrightarrow{\preccurlyeq}$ given by $\gamma_1^\circ \overleftrightarrow{\preccurlyeq} \gamma_2^\circ$ iff $\gamma_2^\circ \preccurlyeq \gamma_1^\circ$.

Although less insightful, it is possible to give an equivalent definition of $\overleftrightarrow{C}$ without referencing the partial order. For any $x \in \cup_{(\gamma^\circ, c) \in C} s(\gamma^\circ)$ let $\check{c}(x) := \max\{c(x)\colon\ (\gamma^\circ, c) \in C\}$ denote the maximal color used for x by a coloring in C. Then, denoting $\overleftrightarrow{c}(x) := \check{c}(x) - c(x)$ for each coloring c, the inverse of C is the set

$$\overleftrightarrow{C} := \{(\gamma^\circ, \overleftrightarrow{c})\colon\ (\gamma^\circ, c) \in C\}.$$

Given a cycle clump C we define the (deterministic) DF-stacks of colored arrows[2] $\{(a_i(x), b_i(x))_{i \in \mathbb{N}_0}\colon\ x \in \mathcal{X}\}$ as follows. If there exists a colored cycle $(\gamma^\circ, \overleftrightarrow{c}) \in \overleftrightarrow{C}$

[2]The arrows $a_i(x)$ are colored either red or green, as indicated by $b_i(x)$, see section 3.1.2.4.

with $\overleftrightarrow{c}(x) = i$, then we set $b_i(x) = 0$ and $a_i(x) = y_1$, where $(y_0, y_1, \dots, y_l) \in \gamma^\circ$ is the unique representative of γ° with $y_0 = x$. Otherwise, if no such colored cycle exists in $\overleftrightarrow{C}$, then we set $b_i(x) = 1$ and $a_i(x) = x$.

So, all green arrows in the stacks correspond to edges that are traversed by cycles in $\overleftrightarrow{C}$.

As is explained in section 3.1.2.4, we can obtain a finite length walk $f(C, v) := (x_0, x_1, \dots, x_l)$ from the vertex r and these DF-stacks, by setting $x_0 := r$ and for $k \in \mathbb{N}_0$ iteratively setting

$$x_{k+1} := a_{i_k}(x_k), \quad \text{with } i_k := \ell[(x_j)_{0 \le j \le k-1}](x_k), \tag{4.16}$$

and $l = \min\{k \in \mathbb{N}_0 \colon \ b_{i_k}(x_k) = 1\}$.

For a closed walk $\gamma = (x_0, \dots, x_l) \in \mathcal{P}^{\mathrm{cl}}$ the inverse $f^{-1}(\gamma)$ can be described using the loop-erasure procedure, and the sequence of self-avoiding walks $(\gamma_i)_{0 \le i \le l}$ produced by this procedure, see section 3.1.2.1.

For each $i \in [l]$ with $x_i \in s(\gamma_{i-1})$, we define the cycle γ_i° to be the equivalence class of the closed walk $(y_{k_i}, y_{k_i+1}, \dots, y_m, x_i)$, which is the cycle that is erased by the loop-erasure procedure in iteration i. The coloring $c_i : s(\gamma_i^\circ) \to \mathbb{N}_0$ is given by $c_i(x) := \sum_{j<i} \mathbf{1}_{s(\gamma_j^\circ)}(x)$, i.e. $c_i(x)$ is equal to the number of times x occurs in the support of the previous cycles. We can then iteratively define the collection of colored cycles C_{i+1} by setting $C_0 = \varnothing$ and

$$C_i := \begin{cases} C_{i-1} & \text{if } x_i \notin s(\gamma_{i-1}) \\ C_{i-1} \cup \{(\gamma_i^\circ,\, c_i)\} & \text{if } x_i \in s(\gamma_{i-1}). \end{cases} \tag{4.17}$$

It then holds that $f^{-1}(\gamma) = \left(\overleftrightarrow{C}_l, x_0\right)$.

By construction it holds that $f^{-1}(f(C, v)) = (C, v)$, which concludes the proof. □

4.5.2 Constructing the Poissonian loop-ensemble

For a set of colored cycles C, we introduce its upward color shift $C_\uparrow$, which is the set of colored cycles defined as

$$C_\uparrow := \{(\gamma^\circ,\, c - \hat{c}|_{s(\gamma^\circ)}) \colon \ (\gamma^\circ,\, c) \in C\}, \tag{4.18}$$

where $\hat{c} : \cup_{(\gamma^\circ,\, c) \in C} \ s(\gamma^\circ) \to \mathbb{N}_0$ is given by $\hat{c}(x) = \min\{c(x) \colon \ (\gamma^\circ,\, c) \in C\}$, and $\hat{c}|_{s(\gamma^\circ)}$ denotes its restriction to the support of γ°.

Note that if C is admissible, then it holds for its up-shift that $C_\uparrow = C$.

Rather than considering the total set of colored cycles produced until time t, we are concerned with the colored cycles produced at a single time point

$$\mathfrak{C}_t^\Delta := \mathfrak{C}_t \setminus \bigcup_{s<t} \mathfrak{C}_s. \tag{4.19}$$

While $\mathfrak{C}_t^\Delta$ is itself not necessarily admissible, its up-shift $(\mathfrak{C}_t^\Delta)_\uparrow$ is. We remark that for the up-shift of $\mathfrak{C}_t^\Delta$ it holds that

$$(\mathfrak{C}_t^\Delta)_\uparrow = \{(\gamma^\circ,\, c - (N_t - \underline{1})|_{s(\gamma^\circ)})\colon\ (\gamma^\circ,\, c) \in \mathfrak{C}_t^\Delta\}.$$

For all jump times $\tau \in \mathcal{T}$ the DF-stacks $\{(A_i(x), U_i(x))_{i\in\mathbb{N}_0}\colon\ x \in \mathcal{X}\}$ can be used to define a vertex V_τ, which will be called the *waking root* at time τ. The waking root denotes the root of the coupled forest process $(\Phi_{1/t})_{t\geq 0}$ at which an arrow in the DF-stack changes its dynamic color from red to green at time τ, i.e. V_τ is the unique vertex v such that $\lim_{s\uparrow\tau} U_{N_s(v)-1}(v) = \frac{1}{1+\delta\tau}$.

Importantly, the waking root V_τ is measurable with respect to the σ-algebra generated by $(\mathfrak{C}_s, \Phi_{1/s})_{0\leq s\leq\tau}$, since V_τ is the only root of the forest $\Phi_{\frac{1}{\tau-\mathrm{d}t}}$ just before time τ at which the occupation field increases. In fact, as will become clear from the definition in eq. (4.22), the Poisson point process $\mathcal{L}$ of theorem 4.1 is measurable with respect to the σ-algebra generated by both $(\mathfrak{C}_t)_{t\geq 0}$ and the collection of waking roots.

Definition 4.5.1. The *cycle clump process* $\mathcal{C}$ is the random atomic measure on the product space $\mathcal{A}\times\mathcal{X}\times(0,\infty)$ defined by

$$\mathcal{C}(\{C\}\times\{x\}\times(a,b]) := |\{\tau\in\mathcal{T}\cap(a,b]\colon\ (\mathfrak{C}_\tau^\Delta)_\uparrow = C,\ V_\tau = x\}|. \tag{4.20}$$

■

An important observation is that the random atomic measure on $\mathcal{A}\times\mathcal{X}$ given by

$$\{C\}\times\{x\}\mapsto \mathcal{C}(\{C\}\times\{x\}\times(0,\infty)) \tag{4.21}$$

is supported on the set of rooted cycle clumps. Together, the above observation and the bijection f from lemma 4.4 allow us to define a random atomic measure $\mathcal{L}$ on $\mathcal{P}^{\mathrm{cl}}\times(0,\infty)$ by

$$\mathcal{L}(\{\gamma\}\times(a,b]) := \mathcal{C}(\{f^{-1}(\gamma)\}\times(a,b]). \tag{4.22}$$

As the notation suggests, $\mathcal{L}$ is indeed the sought Poisson point process of theorem 4.1, as will become clear in the remainder of this section.

For a collection of colored cycles C write $l(C) := \sum_{(\gamma^\circ,\, c)\in C} |e(\gamma^\circ)|$ to denote the total number of traversed edges by all cycles in C.

Lemma 4.5. *The the cycle clump process $\mathcal{C}$ is a Poisson point process on $\mathcal{A}\times\mathcal{X}\times(0,\infty)$ with intensity measure*

$$\{C\}\times\{v\}\times(a,b]\mapsto \mathbf{1}\{(C,v)\in \mathrm{RCC}\}\int_a^b \tfrac{1}{(1+\delta t)^2}\left(\tfrac{t}{1+\delta t}\right)^{l(C)-1}\prod_{(\gamma^\circ,\, c)\in C} w(e(\gamma^\circ))\ \mathrm{d}t,$$

where RCC *denotes the set of rooted cycle clumps.*

Proof. For any admissible set of colored cycles $C\in\mathcal{A}$ and any rooted forest F, it holds by Wilson's formula that

$$\mathbb{P}(\mathfrak{C}_t = C,\ \Phi_{1/t} = F) = \left(\tfrac{1}{1+\delta t}\right)^{r(F)}\left(\tfrac{t}{1+\delta t}\right)^{|F|+l(C)} w(F)\prod_{(\gamma^\circ,\, c)\in C} w(e(\gamma^\circ)), \tag{4.23}$$

where we recall that $r(F)$ and $|F|$ denote the number of roots and edges of F, respectively. In eq. (4.24), we extend this formula to the dynamic setting with DF-stacks $\{(A_i(x), U_i(x))_{i\in\mathbb{N}_0}\colon\ x \in \mathcal{X}\}$ that contain uniform killing marks. For an admissible set of colored cycles C and $x \in \mathcal{X}$ we write

$$c^*_C(x) := \begin{cases} 1 + \max\{c(x)\colon\ (\gamma^\circ,\, c) \in C,\ x \in s(\gamma^\circ)\} & \text{if } x \in \bigcup_{(\gamma^\circ,\, c)\in C} s(\gamma^\circ) \\ 0 & \text{otherwise.} \end{cases}$$

to denote the smallest color that is not used for vertex x by a colored cycle in C.

Let $\mathbf{u}^* : \mathcal{X} \to [0,1)$ be such that $\mathbf{u}^*(x) < \frac{1}{1+\delta t}$ for all $x \in \rho(F)$ and $\mathbf{u}^*(x) \geq \frac{1}{1+\delta t}$ for all $x \notin \rho(F)$. Write $k := |C|$ and $C = \{(\gamma_1^\circ, c_1), (\gamma_2^\circ, c_2), \dots, (\gamma_k^\circ, c_k)\}$, and for all $i \in [k]$ let $\mathbf{u}_i : s(\gamma_1^\circ) \to [\frac{1}{1+\delta t}, 1)$ be given. It holds that

$$\begin{aligned} &\mathbb{P}\left(\bigcap_{i\in[k]} \bigcap_{x\in s(\gamma_i^\circ)} U_{c_i(x)}(x) \in \mathrm{d}\mathbf{u}_i(x),\ \bigcap_{x\in\mathcal{X}} U_{c^*_C(x)}(x) \in \mathrm{d}\mathbf{u}^*(x) \,\middle|\, \mathfrak{C}_t = C,\ \Phi_{1/t} = F \right) \\ &= \left(\prod_{i\in[k]} \prod_{x\in s(\gamma_i^\circ)} \tfrac{1+\delta t}{\delta t}\ \mathrm{d}\mathbf{u}_i(x) \right) \left(\prod_{x\notin\rho(F)} \tfrac{1+\delta t}{\delta t}\ \mathrm{d}\mathbf{u}^*(x) \right) \left(\prod_{x\in\rho(F)} (1+\delta t)\ \mathrm{d}\mathbf{u}^*(x) \right), \end{aligned} \tag{4.24}$$

where, by abuse of notation, $\mathrm{d}\mathbf{u}_i(x)$ denotes both a small enough real number and the interval $(\mathbf{u}_i(x), \mathbf{u}_i(x) + \mathrm{d}\mathbf{u}_i(x)]$.[3]

We will use eq. (4.24) to compute the probability

$$\mathbb{P}\left(\mathcal{C}\left(\{C'\} \times \{v\} \times \mathrm{d}t\right) = 1 \,\middle|\, \mathfrak{C}_{t+\mathrm{d}t} = C \cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}} = F \right), \tag{4.25}$$

where (C', v) is a rooted cycle clump, and we denote by

$$C'_{\downarrow C} := \{(\gamma^\circ,\, c + c^*_C)\colon\ (\gamma^\circ,\, c) \in C'\} \tag{4.26}$$

the set of colored cycles obtained by downwards shifting the colors of cycles in C' to fit underneath the colored cycles in C.

Conditionally on the event that both $\mathfrak{C}_{t+\mathrm{d}t} = C \cup C'_{\downarrow C}$ and $\Phi_{\frac{1}{t+\mathrm{d}t}} = F$, the event that both $\mathfrak{C}_t = C$ and $\mathcal{C}\left(\{C'\} \times \{v\} \times \mathrm{d}t\right) = 1$ can be expressed in terms of the uniform killing marks as the intersection of the following events:

(1) for all $(\gamma^\circ,\, c) \in C$ and all $x \in s(\gamma^\circ)$ it holds that $U_{c(x)}(x) \geq \frac{1}{1+\delta t}$;

(2) $U_{c^*_C(v)}(v) \in [\frac{1}{1+\delta(t+\mathrm{d}t)}, \frac{1}{1+\delta t})$;

(3) for all $x \in s(\gamma^\circ_{\max}) \setminus v$ it holds that $U_{c^*_C(x)}(x) \geq \frac{1}{1+\delta t}$

(4a) for all $(\gamma^\circ,\, c) \in C'_{\downarrow C} \setminus \{(\gamma^\circ_{\max}, c^*_C|_{s(\gamma^\circ_{\max})})\}$ and all $x \in s(\gamma^\circ)$ it holds that $U_{c(x)}(x) > U_{c^*_C(v)}(v)$,

[3] In this context, 'small enough' means that $\mathbf{u}_i(x) + \mathrm{d}\mathbf{u}_i(x) \leq 1$ for $i \in [k]$, $\mathbf{u}^*(x) + \mathrm{d}\mathbf{u}^*(x) \leq 1$ for $x \notin \rho(F)$, and $\mathbf{u}^*(x) + \mathrm{d}\mathbf{u}^*(x) < \frac{1}{1+\delta t}$ for $x \in \rho(F)$.

where we recall that $\gamma^\circ_{\max}$ denotes the maximal cycle in the cycle clump C', and $c^*_C|_{s(\gamma^\circ_{\max})}$ denotes the restriction of c^*_C to the support of γ°, which is the coloring of the maximal colored cycle in $C'_{\downarrow C}$. That is,

$$\begin{aligned}&\mathbb{P}\left(\mathfrak{C}_t = C,\ \mathcal{C}\left(\{C'\}\times\{v\}\times \mathrm{d}t\right)=1 \;\middle|\; \mathfrak{C}_{t+\mathrm{d}t} = C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}} = F\right)\\ &= \mathbb{P}\left((1)\cap(2)\cap(3)\cap(4a) \;\middle|\; \mathfrak{C}_{t+\mathrm{d}t} = C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}} = F\right).\end{aligned} \tag{4.27}$$

By defining the event

(4*b*) for all $(\gamma^\circ,\, c)\in C'_{\downarrow C}\backslash\{(\gamma^\circ_{\max}, c^*_C|_{s(\gamma^\circ_{\max})})\}$ and all $x\in s(\gamma^\circ)$ it holds that $U_{c(x)}(x)\geq \frac{1}{1+\delta t}$,

in the limit as $\mathrm{d}t\downarrow 0$ we have by eq. (4.24) that

$$\begin{aligned}&\mathbb{P}\left(\mathfrak{C}_t = C,\ \mathcal{C}\left(\{C'\}\times\{v\}\times \mathrm{d}t\right)=1 \;\middle|\; \mathfrak{C}_{t+\mathrm{d}t} = C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}} = F\right)\\ &\geq \mathbb{P}\left((1)\cap(2)\cap(3)\cap(4b) \;\middle|\; \mathfrak{C}_{t+\mathrm{d}t} = C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}} = F\right)\\ &= \frac{1+\delta(t+\mathrm{d}t)}{\delta(t+\mathrm{d}t)}\left(\frac{1}{1+\delta t}-\frac{1}{1+\delta(t+\mathrm{d}t)}\right)\left(\frac{1+\delta(t+\mathrm{d}t)}{\delta(t+\mathrm{d}t)}\frac{\delta t}{1+\delta t}\right)^{l(C\cup C'_{\downarrow C})-1}\\ &= \frac{1}{1+\delta t}\mathrm{d}t + o(\mathrm{d}t)\end{aligned}$$

and that

$$\begin{aligned}&\mathbb{P}\left(\mathfrak{C}_t = C,\ \mathcal{C}\left(\{C'\}\times\{v\}\times \mathrm{d}t\right)=1 \;\middle|\; \mathfrak{C}_{t+\mathrm{d}t} = C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}} = F\right)\\ &\leq \mathbb{P}\left((2) \;\middle|\; \mathfrak{C}_{t+\mathrm{d}t} = C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}} = F\right) = \frac{1+\delta(t+\mathrm{d}t)}{\delta(t+\mathrm{d}t)}\left(\frac{1}{1+\delta t}-\frac{1}{1+\delta(t+\mathrm{d}t)}\right)\\ &= \frac{1}{1+\delta t}\mathrm{d}t + o(\mathrm{d}t).\end{aligned}$$

Upper and lower bounding the probability in eq. (4.25) by

$$\begin{aligned}&\mathbb{P}\left(\mathfrak{C}_t = C,\ \mathcal{C}\left(\{C'\}\times\{v\}\times \mathrm{d}t\right)=1 \;\middle|\; \mathfrak{C}_{t+\mathrm{d}t} = C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}} = F\right)\\ &\leq \mathbb{P}\left(\mathcal{C}\left(\{C'\}\times\{v\}\times \mathrm{d}t\right)=1 \;\middle|\; \mathfrak{C}_{t+\mathrm{d}t} = C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}} = F\right)\\ &\leq \mathbb{P}\left(\mathfrak{C}_t = C,\ \mathcal{C}\left(\{C'\}\times\{v\}\times \mathrm{d}t\right)=1 \;\middle|\; \mathfrak{C}_{t+\mathrm{d}t} = C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}} = F\right)\\ &\quad+\mathbb{P}\left(\mathcal{C}\left(\mathcal{A}\times\mathcal{X}\times \mathrm{d}t\right)\geq 2 \;\middle|\; \mathfrak{C}_{t+\mathrm{d}t} = C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}} = F\right)\\ &= \mathbb{P}\left(\mathfrak{C}_t = C,\ \mathcal{C}\left(\{C'\}\times\{v\}\times \mathrm{d}t\right)=1 \;\middle|\; \mathfrak{C}_{t+\mathrm{d}t} = C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}} = F\right)+o(\mathrm{d}t),\end{aligned} \tag{4.28}$$

gives us that

$$\mathbb{P}\left(\mathcal{C}\left(\{C'\}\times\{v\}\times \mathrm{d}t\right)=1 \;\middle|\; \mathfrak{C}_{t+\mathrm{d}t} = C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}} = F\right) = \frac{1}{1+\delta t}\mathrm{d}t + o(\mathrm{d}t).$$

Therefore, by eq. (4.23), summing over all admissible sets of colored cycles and all rooted forests gives us that

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{C}\left(\{C'\}\times\{v\}\times \mathrm{d}t\right)=1\right)\\
&=\sum_{C}\sum_{F}\mathbb{P}\left(\mathcal{C}\left(\{C'\}\times\{v\}\times \mathrm{d}t\right)=1,\ \mathfrak{C}_{t+\mathrm{d}t}=C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}}=F\right)\\
&=\left(\frac{\mathrm{d}t}{1+\delta t}+o(\mathrm{d}t)\right)\sum_{C}\sum_{F}\mathbb{P}(\mathfrak{C}_{t+\mathrm{d}t}=C\cup C'_{\downarrow C},\ \Phi_{\frac{1}{t+\mathrm{d}t}}=F)\\
&=\left(\tfrac{\mathrm{d}t}{1+\delta t}+o(\mathrm{d}t)\right)\left(\tfrac{t+\mathrm{d}t}{1+\delta(t+\mathrm{d}t)}\right)^{l(C')}\left(\prod_{(\gamma^\circ,\,c)\in C'}w(e(\gamma^\circ))\right)\sum_{C}\sum_{F}\mathbb{P}(\mathfrak{C}_{t+\mathrm{d}t}=C,\ \Phi_{\frac{1}{t+\mathrm{d}t}}=F)\\
&=\frac{\mathrm{d}t}{1+\delta t}\left(\frac{t}{1+\delta t}\right)^{l(C')}\left(\prod_{(\gamma^\circ,\,c)\in C'}w(e(\gamma^\circ))\right)+o(\mathrm{d}t).
\end{aligned}\tag{4.29}$$

In a similar manner, we compute for $0<s<t$ and any two rooted cycle clumps (C'_1,v_1) and (C'_2,v_2) the probability

$$\mathbb{P}\left(\mathcal{C}\left(\{C'_1\}\times\{v_1\}\times \mathrm{d}s\right)=1,\ \mathcal{C}\left(\{C'_2\}\times\{v_2\}\times \mathrm{d}t\right)=1\right),$$

to show that $\mathcal{C}\left(\{C'_1\}\times\{v_1\}\times \mathrm{d}s\right)$ and $\mathcal{C}\left(\{C'_2\}\times\{v_2\}\times \mathrm{d}t\right)$ are independent. Let C_1 and C_2 be any two admissible sets of colored cycles. It is notationally convenient to define the following four admissible sets of colored cycles

$$\begin{aligned}
C_s &:= C_1\\
C_{s+\mathrm{d}s} &:= C_s\cup (C'_1)_{\downarrow C_s}\\
C_t &:= C_{s+\mathrm{d}s}\cup (C_2)_{\downarrow C_{s+\mathrm{d}s}}\\
C_{t+\mathrm{d}t} &:= C_t\cup (C'_2)_{\downarrow C_t}.
\end{aligned}$$

As in eq. (4.28), we have that

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{C}\left(\{C'_1\}\times\{v_1\}\times \mathrm{d}s\right)=1,\ \mathcal{C}\left(\{C'_2\}\times\{v_2\}\times \mathrm{d}t\right)=1\right)\\
&=\mathbb{P}\begin{pmatrix}\mathfrak{C}_s=C_s, & \mathcal{C}\left(\{C'_1\}\times\{v_1\}\times \mathrm{d}s\right)=1\\ \mathfrak{C}_t=C_t, & \mathcal{C}\left(\{C'_2\}\times\{v_2\}\times \mathrm{d}t\right)=1\end{pmatrix}+o(\mathrm{d}s)+o(\mathrm{d}t).
\end{aligned}$$

It further holds for any rooted forest F that

$$\begin{aligned}
&\mathbb{P}\left(\begin{matrix}\mathfrak{C}_s=C_s, & \mathcal{C}\left(\{C'_1\}\times\{v_1\}\times \mathrm{d}s\right)=1\\ \mathfrak{C}_t=C_t, & \mathcal{C}\left(\{C'_2\}\times\{v_2\}\times \mathrm{d}t\right)=1\end{matrix}\ \middle|\ \begin{matrix}\mathfrak{C}_{s+\mathrm{d}s}=C_{s+\mathrm{d}s},\\ \mathfrak{C}_{t+\mathrm{d}t}=C_{t+\mathrm{d}t},\end{matrix}\quad \Phi_{\frac{1}{t+\mathrm{d}t}}=F\right)\\
&=\frac{\mathrm{d}s\ \mathrm{d}t}{(1+\delta s)(1+\delta t)}+o(\mathrm{d}s)+o(\mathrm{d}t).
\end{aligned}$$

We therefore have that

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{C}\left(\{C_1'\}\times\{v_1\}\times \mathrm{d}s\right)=1,\ \mathcal{C}\left(\{C_2'\}\times\{v_2\}\times \mathrm{d}t\right)=1\right)\\
&=\left(\tfrac{\mathrm{d}s\ \mathrm{d}t}{(1+\delta s)(1+\delta t)}+o(\mathrm{d}s)+o(\mathrm{d}t)\right)\sum_{C_1,C_2}\sum_{F}\mathbb{P}\left(\mathfrak{C}_{s+\mathrm{d}s}=C_{s+\mathrm{d}s},\mathfrak{C}_{t+\mathrm{d}t}=C_{t+\mathrm{d}t},\ \Phi_{\frac{1}{t+\mathrm{d}t}}=F\right)\\
&=\left(\frac{\mathrm{d}s\ \mathrm{d}t}{(1+\delta s)(1+\delta t)}+o(\mathrm{d}s)+o(\mathrm{d}t)\right)\left(\frac{s+\mathrm{d}s}{1+\delta(s+\mathrm{d}s)}\right)^{l(C_1')}\left(\frac{t+\mathrm{d}t}{1+\delta(t+\mathrm{d}t)}\right)^{l(C_2')}\\
&\quad\times\sum_{C_1,C_2}\sum_{F}\mathbb{P}(\mathfrak{C}_{s+\mathrm{d}s}=C_1,\ \mathfrak{C}_{t+\mathrm{d}t}=C_1\cup(C_2)_{\downarrow C_1},\ \Phi_{\frac{1}{t+\mathrm{d}t}}=F)\\
&=\mathbb{P}\left(\mathcal{C}\left(\{C_1'\}\times\{v_1\}\times \mathrm{d}s\right)=1\right)\mathbb{P}\left(\mathcal{C}\left(\{C_2'\}\times\{v_2\}\times \mathrm{d}t\right)=1\right)+o(\mathrm{d}s)+o(\mathrm{d}t).
\end{aligned}$$

The above computations show that the random measures of any two disjoint subsets are independent. Independence for any k subsets is shown identically.

By the above independence, it follows from Kingman's representation theorem that $\mathcal{C}$ is a Poisson point process [46]. Since the rate function has been computed in eq. (4.29) above, this completes the proof. □

Corollary 4.5.1. *The the random atomic measure $\mathcal{L}$ is a Poisson point process on $\mathcal{P}^{\mathrm{cl}}\times(0,\infty)$ with intensity measure*

$$\{\gamma\}\times(a,b]\mapsto\int_a^b \kappa_t\ \mathbf{P}_{\mathrm{unif}}((X_k)_{0\le k\le T_t+1}=\gamma\mid X_{T_t+1}=X_0)\ \mathrm{d}t.$$

Proof of theorem 4.1. Theorem 4.1 follows from corollary 4.5.1 and the computations in the proof of corollary 3.6.1. □

4.5.3 Spectral decomposition

The result of theorem 4.2 is a direct consequence of corollary 4.5.1.

Proof of theorem 4.2. By corollary 4.5.1 the intensity measure of $\mathcal{L}$ is given by

$$\{\gamma\}\times(a,b]\mapsto\int_a^b \kappa_t\ \mathbf{P}_{\mathrm{unif}}((X_k)_{0\le k\le T_t+1}=\gamma\mid X_{T_t+1}=X_0)\ \mathrm{d}t.$$

Writing $\gamma=(x_0,x_1\ldots,x_l)$ gives us by the independence of X and T_t that

$$\begin{aligned}
&\mathbf{P}_{\mathrm{unif}}((X_k)_{0\le k\le T_t+1}=\gamma\mid X_{T_t+1}=X_0)\\
&\quad=P(\Gamma_{l,x_0}=\gamma)\mathbf{P}_{\mathrm{unif}}(T_t+1=l,\ X_0=x_0\mid X_{T_t+1}=X_0).
\end{aligned}$$

Using eq. (3.57), we have for any $k\in\mathbb{N}$ and $x\in\mathcal{X}$ that

$$\begin{aligned}
&\mathbf{P}_{\mathrm{unif}}(T_t+1=k,\ X_0=x\mid X_{T_t+1}=X_0)\\
&\quad=\frac{\mathbf{P}_{\mathrm{unif}}\left(X_0=x,\ X_k=x\mid T_t+1=k\right)\mathbf{P}_{\mathrm{unif}}(T_t+1=k)}{\mathbf{P}_{\mathrm{unif}}\left(X_{T_t+1}=X_0\right)}\\
&\quad=\frac{\mathbf{P}_{\mathrm{unif}}\left(X_0=x,\ X_k=x\right)}{\frac{1+\delta t}{n\delta}\kappa_t}\left(\tfrac{\delta t}{1+\delta t}\right)^{k-1}\tfrac{1}{1+\delta t}\\
&\quad=\tfrac{1}{t\kappa_t(1+\delta t)}\left(\tfrac{\delta t}{1+\delta t}\right)^{k}\mathbf{P}_x\left(X_k=x\right)=\tfrac{1}{t\kappa_t(1+\delta t)}\left(\tfrac{\delta t}{1+\delta t}\right)^{k}\left(I-\tfrac{1}{\delta}L\right)^k(x,x).
\end{aligned}$$

For two measures ν_f and ν_g with Radon-Nikodym derivatives $f := \frac{d\nu_f}{d\mu}$ and $g := \frac{d\nu_g}{d\mu}$ we write $\langle \nu_f, \nu_g \rangle_\mu^* := \langle f, g \rangle_\mu = \sum_{x \in \mathcal{X}} \frac{\nu_f(x)\nu_g(x)}{\mu(x)}$. Denote by $\nu_0, \ldots, \nu_{n-1}$ the left eigenmeasures of L, so that $\frac{d\nu_j}{d\mu} = \mathbf{v}_j$. Note that each eigenvector $\mathbf{v}_j$ of L is also an eigenvector of $I - \frac{1}{\delta}L$ with eigenvalue $1 - \frac{\lambda_j}{\delta}$. Hence, by letting $\mathbf{1}_x$ denote the indicator of x and δ_x the Dirac-measure at x, we have that

$$\begin{aligned}
\left(I - \tfrac{1}{\delta}L\right)^k (x,x) = \delta_x \left(I - \tfrac{1}{\delta}L\right)^k \mathbf{1}_x &= \left(\sum_{i<n} \langle \nu_i, \delta_x \rangle_\mu^* \, \nu_i\right) \left(\sum_{j<n} \langle \mathbf{v}_j, \mathbf{1}_x \rangle_\mu \left(1 - \tfrac{\lambda_j}{\delta}\right)^k \mathbf{v}_j\right) \\
&= \sum_{i<n} \sum_{j<n} \left(1 - \tfrac{\lambda_j}{\delta}\right)^k \langle \nu_i, \delta_x \rangle_\mu^* \langle \mathbf{v}_j, \mathbf{1}_x \rangle_\mu \, \nu_i \mathbf{v}_j \\
&= \sum_{j<n} \left(1 - \tfrac{\lambda_j}{\delta}\right)^k \langle \nu_j, \delta_x \rangle_\mu^* \langle \mathbf{v}_j, \mathbf{1}_x \rangle_\mu = \frac{1}{\delta^k} \sum_{j<n} (\delta - \lambda_j)^k \, \mathbf{v}_j(x)^2 \mu(x).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathbf{P}_{\text{unif}}(X_0 = x,\ T_t + 1 = k \mid X_{T_t+1} = X_0) \\
&= \frac{1}{t\kappa_t(1+\delta t)} \left(\frac{t}{1+\delta t}\right)^k \sum_{j<n} (\delta - \lambda_j)^k \, \mathbf{v}_j(x)^2 \mu(x) \\
&= \frac{1}{\kappa_t} \sum_{j<n} \frac{\delta - \lambda_j}{(1+\delta t)(1+\lambda_j t)} P(T_j + 1 = k) P(Y_j = x),
\end{aligned}$$

which concludes the proof. □

PART II

COUPLINGS AND MATCHINGS

CHAPTER 5
Couplings and Matchings

This chapter is based on the following paper: T. Koperberg. "Couplings and Matchings: Combinatorial notes on Strassen's theorem". In: *Statistics & Probability Letters* 209 (2024), p. 110089.

Abstract

Some mathematical theorems represent ideas that are discovered again and again in different forms. One such theorem is Hall's marriage theorem. This theorem is equivalent to several other theorems in combinatorics and optimization theory, in the sense that these results can easily be derived from each other. Remarkably, this equivalence extends to Strassen's theorem, a celebrated result on couplings of probability measures.

In this paper the equivalence between Strassen's theorem and Hall's theorem is investigated from a combinatorial perspective. A novel combinatorial lemma will be introduced that can be used to deduce both Hall's theorem and a finite version of Strassen's theorem, providing a simple proof of their equivalence.

5.1 Introduction

In the original paper from 1935 Hall already mentions a similarity between his *marriage theorem* and a result by König from 1916. Since then numerous other results have been found that are 'equivalent' to Hall's theorem. This equivalence is an informal concept and simply means that two results can be derived from each other via simple proofs. This class of equivalent theorems includes among others *Menger's theorem* (1927), *König's minimax theorem* (1931), the *Birkhoff-von Neumann theorem* (1946), *Dilworth's theorem* (1950) and the *max-flow min-cut theorem* by Ford and Fulkerson (1956). An extensive discussion on these equivalences can be found in [73].

Another result that belongs to this class of equivalent statements is *Strassen's theorem* (1965). As this theorem is a result from probability theory, the original proof made use of analytical tools rather than the combinatorial methods used in the proofs of the above mentioned theorems. Therefore, it is remarkable that this result is, in fact, equivalent to these combinatorial statements. This relation between Strassen's theorem and Hall's theorem is already known in the literature, as Dudley used Hall's marriage theorem to prove an extension of Strassen's original result [23].

In this paper we will look at Strassen's theorem from a combinatorial perspective. Our discussion will be restricted to a finite variant of Strassen's theorem, which is stated in theorem 5.1. For the general version of the theorem the reader is referred to [57].

The goal of this paper is twofold: firstly to give a combinatorial proof of the finite version of Strassen's theorem directly from first principles, and secondly to give a simple proof of the equivalence between Strassen's theorem and Hall's theorem, which will be done using an adaptation of Dudley's proof. For both of these objectives we will make use of a novel lemma, that will be introduced in section 5.2.1, and which will be referred to as the *subforest lemma*. As will be discussed in remark 5.4, this lemma could be derived from a more abstract result within the theory of optimal transport. In section 5.1.1 we will introduce the two main theorems. The content of the subsequent sections is outlined in fig. 5.1.

5.1.1 The two main theorems

We will start by introducing the two theorems that are the main topic of this paper.

If $\mathbb{P}$ and $\mathbb{P}'$ are probability measures on two finite sets A and B, respectively, then a *coupling* of $\mathbb{P}$ and $\mathbb{P}'$ is any probability measure $\hat{\mathbb{P}}$ on the product set $A \times B$ for which its marginals correspond to $\mathbb{P}$ and $\mathbb{P}'$. That is, for all $U \subseteq A$ and $S \subseteq B$ it holds that $\mathbb{P}(U) = \hat{\mathbb{P}}(U \times B)$ and $\mathbb{P}'(S) = \hat{\mathbb{P}}(A \times S)$.

Theorem 5.1 (Strassen's theorem for finite sets). *Let A and B be finite sets and $R \subseteq A \times B$ a relation between them. Let $\mathbb{P}$ and $\mathbb{P}'$ be probability measures on A and B, respectively. Then there exists a coupling $\hat{\mathbb{P}}$ of $\mathbb{P}$ and $\mathbb{P}'$ with $\hat{\mathbb{P}}(R) = 1$ if and only if*

$$\mathbb{P}(U) \leq \mathbb{P}'(N_R(U)), \quad \text{for all } U \subseteq A, \tag{5.1}$$

where $N_R(U) = \{y \in B \colon \exists x \in U \text{ s.t. } (x, y) \in R\}$.

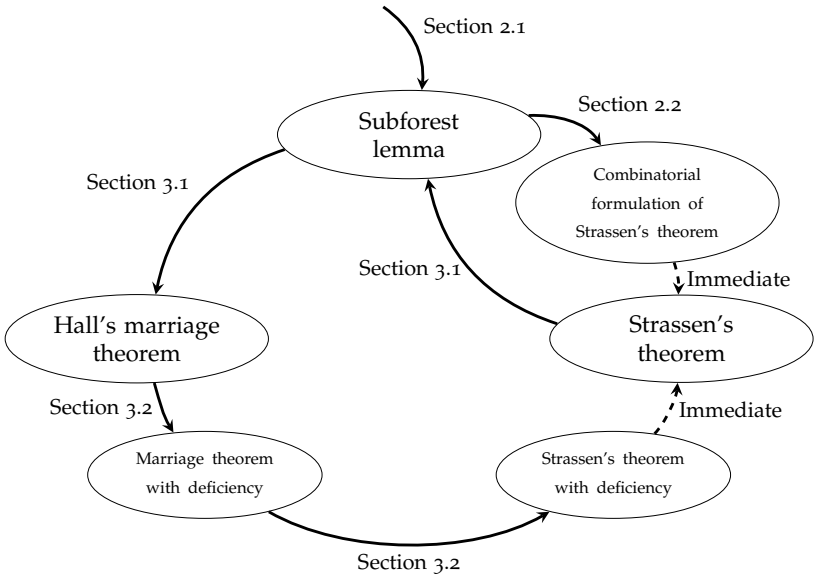


Figure 5.1: A graphical outline of this paper, where the arrows represent the different proofs.

We will refer to (5.1) as the *coupling condition*. In [26] it is shown how the general version of Strassen's theorem can be derived from this finite version.

In this paper we will use the graph theoretic formulation of the marriage theorem. All graphs in this paper are assumed to be simple finite undirected graphs. A *bipartite graph* is a graph of which the vertices can be partitioned into two sets $\{A, B\}$ such that all edges have one endpoint in A and the other endpoint in B. This partition $\{A, B\}$ will be called the *bipartition* of the graph. A *matching* of a graph is a subset M of its edges such that all vertices are incident to at most one edge in M. If all vertices are incident to an edge in M, then M is called a *perfect matching*.

Theorem 5.2 (Hall's marriage theorem). *Let G be a bipartite graph with bipartition $\{A, B\}$ such that $|A| = |B|$. Then G contains a perfect matching if and only if it holds that*

$$|U| \leq |N_G(U)|, \qquad \text{for all } U \subseteq A. \tag{5.2}$$

Here $N_G(U)$ denotes the set of vertices that are neighbors of vertices in U. If the underlying graph is clear, then the subscript will be dropped. We will refer to (5.2) as the *marriage condition*.

5.2 Independent proof of Strassen's theorem

5.2.1 The subforest lemma

A graph that does not contain any cycles is called a *forest*. A *weighted graph* is a graph that is equipped with a vertex weight function $w : V \to [0, \infty)$. For such weight

functions we write $w(U) = \sum_{x \in U} w(x)$ for $U \subseteq V$. Unless otherwise specified, a subgraph of a weighted graph is equipped with the restriction of the weight function to the vertices of the subgraph. So, in particular a spanning subgraph has the same weight function as the underlying full graph. For brevity we will call a spanning subgraph a *subforest* when it is a forest.

Lemma 5.3 (Subforest lemma). *Let $G = (V, E, w)$ be a weighted bipartite graph with bipartition $\{A, B\}$ such that $w(A) = w(B)$. If it holds that*

$$w(U) \leq w(N_G(U)), \quad \text{for all } U \subseteq A, \tag{5.3}$$

then G contains a subforest that satisfies (5.3)*.*

We will refer to (5.3) as the *subforest condition*. Both the marriage condition (5.2) and the coupling condition (5.1) are special cases of this subforest condition. For the marriage condition all vertices have unit weight, while for the coupling condition the weight function is normalized so that $w(A) = w(B) = 1$. Note that these three conditions seem to break the symmetry between sets A and B that is present in the setting of the theorems. This is in fact not the case, as it can be easily verified that (5.3) implies that $w(U) \leq w(N_G(U))$ for all $U \subseteq B$.

Here we give an independent proof of lemma 5.3 directly from first principles. The proof uses the same strategy used in the inductive proof of the marriage theorem by Halmos and Vaughan [34], in which the induction hypothesis acts as a marriage broker. That is, we distinguish between the case where the graph contains a 'critical set' of vertices and the case where no such set exists.

Proof of lemma 5.3. We will apply induction on $|V|$. If $|V| = 2$, then G is itself a forest, so there is nothing to prove. Now assume that $|V| > 2$ and that the statement holds when $|V|$ is smaller. Let $\mathcal{S} = \{U \subseteq A\colon\ 0 < |U| < |A|\} \cup \{U \subseteq B\colon\ 0 < |U| < |B|\}$ denote the collection of non-empty strict subsets of either A or B. We will distinguish two cases.

In the first case we assume that there exists a $U \in \mathcal{S}$ with $w(U) = w(N_G(U))$. Without loss of generality we assume that $U \subseteq A$. Let $\{V_1, V_2\}$ be the partition of V given by $V_1 = U \cup N_G(U)$ and $V_2 = (A \setminus U) \cup (B \setminus N_G(U))$. Then both induced subgraphs $G[V_1]$ and $G[V_2]$ satisfy the subforest condition (5.3). Thus by the induction hypothesis there exist subforests F_1 and F_2 of $G[V_1]$ and $G[V_2]$, respectively, that both satisfy the subforest condition (5.3). The graph $F = (V, E(F_1) \cup E(F_2))$, that contains all edges of F_1 and F_2, is a subforest of G, that satisfies the subforest condition with respect to w. Also note that F contains at least two connected components, since none of the vertices in V_1 is connected to any of the vertices in V_2.

For the second case we assume that $w(U) < w(N_G(U))$ for all $U \in \mathcal{S}$. Let ε denote the minimal weight of any vertex of G, i.e. $\varepsilon = \min_{v \in V} w(v)$. Let $x \in V$ be any vertex with $w(x) = \varepsilon$. Without loss of generality we can assume that $x \in A$. Let $y \in N_G(x)$ be any neighbor of x. Let $\mathcal{U} = \{U \subseteq A\colon\ x \notin U \text{ and } U \cap N_G(y) \neq \varnothing\}$ and take

$$\delta = \min_{U \in \mathcal{U}} w(N_G(U)) - w(U).$$

Since $w(N_G(y)) > w(y) \geq w(x)$, vertex y has at least two neighbors. It follows that $A - x \in \mathcal{U}$ and thus that

$$\delta \leq w(N_G(A-x)) - w(A-x) = w(B) - w(A-x) = \varepsilon.$$

Let $D \in \mathcal{U}$ be such that $w(N_G(D)) - w(D) = \delta$. We add a new element $\tilde{x}$ to A to obtain $\tilde{A} = A + \tilde{x}$. Let $\tilde{V} = V + \tilde{x}$, $\tilde{E} = E(G) + \{\tilde{x}, y\}$ and $\tilde{G} = (\tilde{V}, \tilde{E})$. Define the weight function $\tilde{w}$ on $V + \tilde{x}$ by $\tilde{w} = w + \delta \mathbf{1}_{\{\tilde{x}\}} - \delta \mathbf{1}_{\{x\}}$. The weighted graph $(\tilde{G}, \tilde{w})$ now satisfies the subforest condition, since for all $U \subseteq A - x$ with $y \in N_G(U)$ it holds that $U \in \mathcal{U}$, so that

$$\tilde{w}(U + \tilde{x}) = w(U) + \delta \leq w(N_G(U)) = \tilde{w}(N_{\tilde{G}}(U + \tilde{x})),$$

while if $y \notin N_G(U)$, then

$$\tilde{w}(U + \tilde{x}) \leq w(N_G(U)) + \delta = \tilde{w}(N_{\tilde{G}}(U + \tilde{x})) - \tilde{w}(y) + \delta \leq \tilde{w}(N_{\tilde{G}}(U + \tilde{x})).$$

For subsets $U \subseteq A$ that include x or that do not include $\tilde{x}$ condition (5.3) is also easily verified. We now have that $\tilde{w}(D + \tilde{x}) = \tilde{w}(N_{\tilde{G}}(D + \tilde{x}))$, so that $(\tilde{G}, \tilde{w})$ satisfies the assumptions of the first case.

If $D \neq A - x$, then $|(D + \tilde{x}) \cup N_{\tilde{G}}(D + \tilde{x})| < |V|$. It follows from the induction hypothesis, in the same manner as in the previous case, that there exists a subforest $\tilde{F}$ of $\tilde{G}$ with x and y in two distinct components such that $(\tilde{F}, \tilde{w})$ satisfies (5.3). The graph $F = (V, E(\tilde{F}) - \{\tilde{x}, y\} + \{x, y\})$ is a spanning subgraph of G. Since x and y are contained in distinct components of $\tilde{F}$, we also have that F is a forest. It is also clear that (F, w) satisfies the subforest condition.

If instead we have that $D = A - x$, then $\varepsilon = \delta$. Define the weight function w' on $V - x$ by $w' = w - \delta \mathbf{1}_{\{y\}}$. Then the weighted graph $(G[V - x], w')$ satisfies the subforest condition, since for all $U \subseteq A - x$ with $y \in N_G(U)$ it holds that $U \in \mathcal{U}$, so that

$$w'(U) = w(U) \leq w(N_G(U)) - \delta = w'(N_{G[V-x]}(U)),$$

while if $y \notin N_G(U)$, then

$$w'(U) = w(U) \leq w(N_G(U)) = w'(N_{G[V-x]}(U)).$$

Hence, by the induction hypothesis, there exists a spanning subforest F' of $G[V - x]$ satisfying the subforest condition. Let $F = (V, E(F') + \{x, y\})$. Then F is a subforest of G satisfying the subforest condition.

In both cases we have shown the existence of a spanning subforest that satisfies the subforest condition, thus completing the proof. □

Remark 5.4. The problem of finding a coupling that satisfies the coupling condition (5.1) can also be phrased as an optimal transport problem. A solution to such a transportation problem corresponds to a bipartite graph with weights assigned to the edges. Klee and Witzgall [48] showed that the polytope of feasible solutions has at its vertices exactly those solutions whose accompanying bipartite graph corresponds to a forest. Hence, the subforest lemma can also be derived from this result of Klee and Witzgall.

5.2.2 Deriving Strassen's theorem from the subforest lemma

For the independent proof of Strassen's theorem for finite sets, we show how it can easily be derived from the subforest lemma. It is natural to translate the setting of theorem 5.1 to a weighted bipartite graph $G = (V, E, w)$ defined by

$$V = A \cup B, \quad E = \{\{x, y\} \colon\ (x, y) \in R\}, \text{ and}$$
$$w(x) = \begin{cases} \mathbb{P}(x) & \text{if } x \in A \\ \mathbb{P}'(x) & \text{if } x \in B \end{cases} \tag{5.4}$$

(Here we assume w.l.o.g. that $A \cap B = \varnothing$.) The coupling condition then translates to $w(U) \leq w(N_G(U))$ for all $U \subseteq A$, while the sought coupling becomes an edge weight function $\hat{w} : E \to [0, \infty)$ that satisfies $w(x) = \sum_{e \sim x} \hat{w}(e)$ for all $x \in A$, where the sum is taken over all edges incident to x. Note that the vertex weight function w obtained in this manner will always satisfy $w(A) = w(B) = 1$, due to the fact that probability measures have total mass 1. This translation, which is further illustrated in fig. 5.2, gives us the following equivalent formulation of theorem 5.1, which resembles a weighted version of Hall's marriage theorem. Theorem 5.1 follows directly from proposition 5.5 by normalizing the vertex and edge weights, so that these form probability measures.

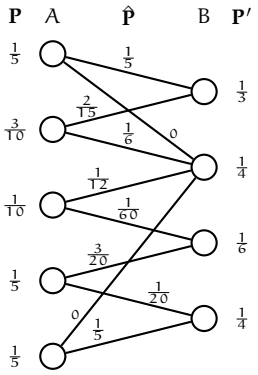


Figure 5.2: A graphical representation of the setting of theorem 5.1. The vertices represent the elements of sets A and B, and the edges represent the elements of R. The number next to a vertex denotes the mass that either $\mathbb{P}$ or $\mathbb{P}'$ assigns to that vertex. The numbers above the edges are the masses assigned by the coupling $\hat{\mathbb{P}}$, which is supported on a subset of R. By interpreting the masses of $\mathbb{P}$ and $\mathbb{P}'$ as vertex weights and the masses of $\hat{\mathbb{P}}$ as edge weights, we find ourselves in the setting of proposition 5.5.

Proposition 5.5 (Combinatorial formulation of Strassen's theorem). *Let $G = (V, E, w)$ be a weighted bipartite graph with bipartition $\{A, B\}$ such that $w(A) = w(B)$. Then the following are equivalent:*

(i) for all $U \subseteq A$ it holds that $w(U) \leq w(N(U))$;

(ii) there exists an edge weight function $\hat{w} : E \to [0, \infty)$ such that for all $x \in V$ it holds that $w(x) = \sum_{e \sim x} \hat{w}(e)$, where the sum is taken over all edges incident to x.

Proof of proposition 5.5 using lemma 5.3. The implication from $((ii))$ to $((i))$ is easily shown. The set of edges with one endpoint in U is a subset of the set of edges with one endpoint in $N(U)$. Hence, if $\hat{w}$ satisfies $((ii))$, then

$$w(U) = \sum_{x \in U} \sum_{e \sim x} \hat{w}(e) \leq \sum_{y \in N(U)} \sum_{e \sim y} \hat{w}(e) = w(N(U)).$$

The reverse implication will be proven by induction on $|V|$. Since w satisfies $((i))$, by the subforest lemma there exists a subforest F of G satisfying $((i))$. Since F is a forest, there exists a vertex x in F with degree 1. Without loss of generality we can assume that $x \in A$. Let $y \in B$ be the unique neighbor of x in F.

Note that it follows from $((i))$ that $w(x) \leq w(y)$. Set $\varepsilon = w(x)$. Consider the induced subgraph $F[V - x]$ obtained by removing vertex x from F and equip it with the vertex weight function $\tilde{w} : V - x \to [0, \infty)$ given by $\tilde{w}(v) = w(v) - \varepsilon \mathbf{1}_{\{v=y\}}$. The weighted graph $(F[V - x], \tilde{w})$ satisfies $((i))$, since for all $U \subseteq A - x$ with $y \in N_F(U)$ it holds that

$$\tilde{w}(U) = w(U) = w(U + x) - \varepsilon \leq w(N_F(U + x)) - \varepsilon = w(N_F(U)) - \varepsilon = \tilde{w}(N_{\tilde{F}}(U)),$$

while for all $U \subseteq A - x$ with $y \notin N_F(U)$ it holds that

$$\tilde{w}(U) = w(U) \leq w(N_F(U)) = \tilde{w}(N_{\tilde{F}}(U)).$$

Hence, by the induction hypothesis there exists an edge weight function $\hat{w}$ on $F[V - x]$ satisfying $((ii))$. Now define an edge weight function on the edges of G by

$$\check{w}(e) = \begin{cases} \hat{w}(e) & \text{if } e \in E(F[V - x]) \\ \varepsilon & \text{if } e = \{x, y\} \\ 0 & \text{otherwise.} \end{cases}$$

Then $\check{w}$ is the sought edge weight function satisfying $((ii))$. □

This independent proof of Strassen's theorem for finite sets is constructive and can in principle be used to find the required coupling. However, far more efficient methods for finding such a coupling exists. In [58, Corollary 2.1.5] it is mentioned that proposition 5.5 can be derived from the max-flow min-cut theorem, see also [60, Theorem 10.4]. The method is similar to the derivation of the marriage theorem from the max-flow min cut theorem, that is given in [27]. This derivation is not only elegant, it also shows that any method for finding maximal network flows can also be used to find such a coupling.

5.3 Equivalence of Hall's theorem and Strassen's theorem

In the second part of this paper we prove the equivalence of Strassen's theorem for finite sets and Hall's marriage theorem.

5.3.1 Deriving Hall's theorem from Strassen's theorem

The derivation of Hall's theorem from Strassen's theorem will go via the subforest lemma.

Proof of lemma 5.3 using Strassen's theorem. The statement will be proven by induction on $|E|$. If $|E| = 1$, then G is itself a forest. Now assume that $|E| \geq 2$ and that the statement holds when $|E|$ is smaller.

Define the probability measures $\mathbb{P}$ and $\mathbb{P}'$ on A and B, respectively, by setting $\mathbb{P}(x) = \frac{w(x)}{w(A)}$ and $\mathbb{P}(y) = \frac{w(y)}{w(B)}$ for $x \in A$ and y in B. Since G satisfies the subforest condition, these two probability measures then satisfy the coupling condition with respect to the relation E. Hence, by Strassen's theorem there exists a coupling $\hat{\mathbb{P}}$ of $\mathbb{P}$ and $\mathbb{P}'$ that is supported on E.

If G is not a forest, then there is a subset of edges $C \subseteq E$ that constitute a cycle. Now take $\varepsilon = \min\{\hat{\mathbb{P}}(e) \colon\ e \in C\}$ and let $e^* \in C$ be such that $\hat{\mathbb{P}}(e^*) = \varepsilon$. Since G is a bipartite graph, the cycle C contains an even number of edges. Hence, we can partition C into two sets $\{I_C, J_C\}$ such that edges in I_C are only incident to edges in J_C and vice-versa. Without loss of generality we can assume that $e^* \in I_C$.

Now define a new probability measure $\tilde{\mathbb{P}}$ on E by

$$\tilde{\mathbb{P}}(e) = \begin{cases} \hat{\mathbb{P}}(e) - \varepsilon & \text{if } e \in I_C \\ \hat{\mathbb{P}}(e) + \varepsilon & \text{if } e \in J_C \\ \hat{\mathbb{P}}(e) & \text{otherwise.} \end{cases}$$

Since each vertex in G is incident to the same number of edges in I_C as to edges in J_C, we have that $\tilde{\mathbb{P}}$ is also a coupling of $\mathbb{P}$ and $\mathbb{P}'$. Moreover, the coupling $\tilde{\mathbb{P}}$ is supported on $E \setminus \{e^*\}$, since by construction it holds that $\tilde{\mathbb{P}}(e^*) = 0$. Thus, by applying Strassen's theorem in the reverse direction, we find that the relation $E \setminus \{e^*\}$ satisfies the coupling condition with respect to $\mathbb{P}$ and $\mathbb{P}'$. It follows that the weighted graph $G - e^*$ satisfies the subforest condition. By the induction hypothesis $G - e^*$ contains a subforest F satisfying that condition. Clearly, F is also a subforest of G, which finishes the proof. □

Proof of theorem 5.2 using the subforest lemma. We will only prove the sufficiency of the marriage condition, which will be done by induction on $|E|$. So, we assume that G satisfies the marriage condition.

Clearly the statement holds if $|E| = 1$. Now assume that $|E| \geq 2$ and that the statement holds if $|E|$ is smaller. Note that the marriage condition is a special case of the subforest condition where each vertex has unit weight. Hence, by lemma 5.3 there exists a subforest F of G that satisfies the marriage condition. Since F is a forest, there exists a vertex x in F with degree 1. Let y be the unique neighbor of x in F. Then the induced subgraph $G[V \setminus \{x, y\}]$ still satisfies the marriage condition. Thus by the induction hypothesis $G[V \setminus \{x, y\}]$ has perfect matching M. Taking $M \cup \{x, y\}$ gives a perfect matching of G. □

5.3.2 Deriving Strassen's theorem from the marriage theorem

To finish our reciprocal derivations we still have to prove Strassen's theorem from the marriage theorem. This will be done using two well-known generalizations of both theorems, propositions 5.6 and 5.7 below, that allow for some small deficiencies in the conditions.

The used generalization of Hall's marriage theorem is due to Ore [66] and can be found in e.g. [58, Thm. 1.3.1]. It can be easily derived from the marriage theorem itself, which led Mirsky to call the marriage theorem a *self-refining result* [65]. For completeness we also give this derivation.

Proposition 5.6 (Hall's theorem with deficiency). *Let G be a bipartite graph with bipartition $\{A, B\}$ with $|A| = |B| = n$. Then G contains a matching M with $|M| \geq n - k$ if and only if it holds that*

$$|U| \leq |N_G(U)| + k, \quad \text{for all } U \subseteq A. \tag{5.5}$$

Proof of proposition 5.6 using Hall's marriage theorem. We only prove the sufficiency of (5.5).

Construct the bipartite graph $\tilde{G}$ by adding k new vertices $a_1, \ldots, a_k$ to A and k new vertices $b_1, \ldots, b_k$ to B. Set $\tilde{A} = A \cup \{a_1, \ldots, a_k\}$ and $\tilde{B} = B \cup \{b_1, \ldots, b_k\}$. Also add edges between all a_i and all vertices in $\tilde{B}$ and all b_i and all vertices in $\tilde{A}$. That is, $\tilde{G} = (\tilde{A} \cup \tilde{B}, \tilde{E})$ with $\tilde{E} = E \cup \{\{a_i, v\} \colon 1 \leq i \leq k,\ v \in \tilde{B}\} \cup \{\{b_i, v\} \colon 1 \leq i \leq k,\ v \in \tilde{A}\}$.

Since G satisfies (5.5) and all vertices have k more neighbors in $\tilde{G}$ than in G, we have that $\tilde{G}$ satisfies the marriage condition (5.2). Hence, by theorem 5.2 $\tilde{G}$ contains a perfect matching $\tilde{M}$. Note that $|\tilde{M}| = n + k$. Hence, $M := \tilde{M} \cap E$ is a matching of G with $|M| \geq n - k$, since at most $2k$ edges in $\tilde{M}$ are incident to any of the $2k$ vertices that are not in G. □

As with proposition 5.6 the generalized version of Strassen's theorem also follows from its original. However, for our purposes we will derive it from proposition 5.6 instead. The proof is an adaptation of the proof by Dudley given in [24, Theorem 11.6.3] adjusted to our finite setting.

Proposition 5.7 (Strassen's theorem with deficiency). *Let A and B be finite sets and $R \subseteq A \times B$ a relation between them. Let $\mathbb{P}$ and $\mathbb{P}'$ be probability measures on A and B, respectively. Let $\varepsilon \geq 0$ be given. Then there exists a coupling $\hat{\mathbb{P}}$ of $\mathbb{P}$ and $\mathbb{P}'$ with $\hat{\mathbb{P}}(R) \geq 1 - \varepsilon$ if and only if*

$$\mathbb{P}(U) \leq \mathbb{P}'(N_R(U)) + \varepsilon, \quad \text{for all } U \subseteq A. \tag{5.6}$$

Proof of proposition 5.7 using proposition 5.6. For the necessity of (5.6) we note that if $\hat{\mathbb{P}}$ is a coupling of $\mathbb{P}$ and $\mathbb{P}'$ with $\hat{\mathbb{P}}(R) \geq 1 - \varepsilon$, then it holds for all $U \subseteq A$ that

$$\mathbb{P}(U) = \hat{\mathbb{P}}(U \times B) \leq \hat{\mathbb{P}}(U \times N_R(U)) + \varepsilon \leq \hat{\mathbb{P}}(A \times N_R(U)) + \varepsilon = \mathbb{P}'(N_R(U)) + \varepsilon.$$

It remains to prove its sufficiency. This will be done in two steps. In the first step we assume that $\mathbb{P}$ and $\mathbb{P}'$ are both rational valued and that $\varepsilon \in \mathbb{Q}$, and in the second step we derive the result for arbitrary $\mathbb{P}$, $\mathbb{P}'$ and ε.

Step (1) Assume that $\mathbb{P}$ and $\mathbb{P}'$ are rational valued and also take ε rational. Define $G = (V, E, w)$ as in (5.4). Then we have that $w(x) \in \mathbb{Q}$ for all $x \in V$. Since V is finite, there exists a large enough $N \in \mathbb{N}$ such that the product $Nw(x)$ is an integer for all $x \in V$ and such that $k := \varepsilon N$ is an integer as well.

Let $\tilde{V} = \bigcup_{x \in A} \bigcup_{i=1}^{Nw(x)} \{x_i\}$ be the set consisting of $Nw(x)$ copies of each element $x \in V$. Now consider the bipartite graph $\tilde{G} = (\tilde{V}, \tilde{E})$, where the edge set is given by $\tilde{E} = \{\{x_i, y_j\} \colon \{x, y\} \in E\}$. That is, two vertices x_i and y_j in $\tilde{G}$ are connected by an edge if and only if their originals x and y are adjacent in G. Denote the bipartition of $\tilde{G}$ by $\{\tilde{A}, \tilde{B}\}$.

Since $\mathbb{P}$ and $\mathbb{P}'$ satisfy (5.6), we then have for all $\tilde{U} \subseteq \tilde{A}$ that

$$|\tilde{U}| \leq |N_{\tilde{G}}(\tilde{U})| + k.$$

By proposition 5.6 there exists a matching $\tilde{M}$ of $\tilde{G}$ with $|\tilde{M}| = N - k$. Let $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$ denote the k vertices in $\tilde{A}$ and $\tilde{B}$, respectively, that are unmatched by $\tilde{M}$. Now consider the set of edges $M^+ := \tilde{M} \cup \{\{a_i, b_i\} \colon i \in [k]\}$, which is obtained from $\tilde{M}$ by adding k arbitrary edges, not necessarily belonging to $\tilde{E}$, between the unmatched vertices.

For each pair $(x, y) \in A \times B$ let

$$\hat{w}(x, y) = \left| \bigcup_{i=1}^{Nw(x)} \bigcup_{j=1}^{Nw(y)} \{x_i, y_j\} \cap M^+ \right|$$

denote the number of edges between copies of x and copies of y that occur in M^+. Since M^+ is a perfect matching of the complete bipartite graph on $\tilde{A} \cup \tilde{B}$, we find that $\sum_{y \in B} \hat{w}(x, y) = Nw(x)$ for all $x \in A$ and similarly that $\sum_{x \in A} \hat{w}(x, y) = Nw(y)$ for all $y \in B$. So, the probability measure $\hat{\mathbb{P}}$ on $A \times B$ defined by $\hat{\mathbb{P}}(x, y) = \frac{\hat{w}(x,y)}{N}$ is a coupling of $\mathbb{P}$ and $\mathbb{P}'$. Since only k of the edges of M^+ do not belong to $\tilde{M}$ we also find that $\hat{\mathbb{P}}(R) = 1 - \varepsilon$, so $\hat{\mathbb{P}}$ is the sought coupling.

Step (2) Let $\mathbb{P}$, $\mathbb{P}'$ and ε be arbitrary. Let $(\varepsilon_i)_{i \in \mathbb{N}}$ be a rational sequence converging to ε from above. Since $\mathbb{P}$ and $\mathbb{P}'$ satisfy (5.6), we can find two sequences $(\mathbb{P}_i)_{i \in \mathbb{N}}$ and $(\mathbb{P}'_i)_{i \in \mathbb{N}}$ of rational valued probability measures that converge to $\mathbb{P}$ and $\mathbb{P}'$, respectively, such that for every $i \in \mathbb{N}$ it holds that

$$\mathbb{P}_i(U) \leq \mathbb{P}'_i(N_R(U)) + \varepsilon_i, \quad \text{for all } U \subseteq A.$$

By the first step of the proof, for each i there exists a coupling $\hat{\mathbb{P}}_i$ of $\mathbb{P}_i$ and $\mathbb{P}'_i$ with $\hat{\mathbb{P}}_i(R) \geq 1 - \varepsilon_i$. We can interpret $(\hat{\mathbb{P}}_i)_{i \in \mathbb{N}}$ as a sequence in the compact metric space $[0, 1]^{|E|}$. Thus it contains a converging subsequence $(\hat{\mathbb{P}}_{i_j})_{j \in \mathbb{N}}$ with limit $\hat{\mathbb{P}}$. It follows that $\hat{\mathbb{P}}(R) \geq 1 - \varepsilon$.

It remains to be shown that $\hat{\mathbb{P}}$ is a coupling of $\mathbb{P}$ and $\mathbb{P}'$. Let $\delta > 0$ be given. Then for all $x \in A$ there exists a $k \in \mathbb{N}$ such that for all $j \geq k$ it holds that both $|\mathbb{P}_{i_j}(x) - \mathbb{P}(x)| < \delta$ and

$$|\hat{\mathbb{P}}_{i_j}(\{x\} \times B) - \hat{\mathbb{P}}(\{x\} \times B)| < \delta.$$

It follows that

$$\begin{aligned}|\mathbb{P}(x) - \hat{\mathbb{P}}(\{x\} \times B)| &< |\mathbb{P}(x) - \hat{\mathbb{P}}_{i_k}(\{x\} \times B)| + \delta \\ &= |\mathbb{P}(x) - \mathbb{P}_{i_k}(x)| + \delta \\ &< 2\delta.\end{aligned}$$

Similarly, we find that $|\mathbb{P}'(x) - \hat{\mathbb{P}}(A \times \{x\})| < 2\delta$ for all $x \in B$. As this holds for all $\delta > 0$, it follows that $\hat{\mathbb{P}}$ is a coupling of $\mathbb{P}$ and $\mathbb{P}'$. □

Bibliography

[1] L. Avena, F. Castell, A. Gaudillière, and C. Mélot. "Random Forests and Networks Analysis". In: *Journal of Statistical Physics* 173 (2018), 985–1027.

[2] L. Avena, F. Castell, A. Gaudillière, and C. Mélot. "Intertwining wavelets or multiresolution analysis on graphs through random forests". In: *Applied and Computational Harmonic Analysis* 48.3 (2020), pp. 949–992.

[3] L. Avena, F. Castell, A. Gaudillière, and C. Mélot. "Approximate and exact solutions of intertwining equations through random spanning forests". In: *In and Out of Equilibrium 3. Celebrating Vladas Sidoravicius*. Vol. 77. Progress in Probability. Springer International Publishing, 2021, pp. 27–69.

[4] L. Avena, J. E. P. Driessen, and V. T. Koperberg. "Loop-erased partitioning via parametric spanning trees: Monotonicities & 1D-scaling". In: *Stochastic processes and their applications* 176 (2024), p. 104436.

[5] L. Avena and A. Gaudillière. "A proof of the transfer-current theorem in absence of reversibility". In: *Statistics & Probability Letters* 142 (2018), pp. 17–22.

[6] L. Avena and A. Gaudillière. "Two Applications of Random Spanning Forests". In: *Journal of Theoretical Probability* 31.4 (2018), pp. 1975–2004.

[7] L. Avena, A. Gaudillière, P. Milanesi, and M. Quattropani. "Loop-erased partitioning of a graph: mean-field analysis". In: *Electronic Journal of Probability* 27 (2022), pp. 1–35.

[8] K. Avrachenkov, P. Chebotarev, and A. Mishenin. "Semi-supervised learning with regularized Laplacian". In: *Optimization Methods & Software* 32.2 (2017), pp. 222–236.

[9] S. Barthelmé, F. Castell, A. Gaudillière, C. Mélot, M. Quattropani, and N. Tremblay. *Spectrum Estimation through Kirchhoff Random Forests*. 2025. arXiv: 2507.19164. URL: https://arxiv.org/abs/2507.19164.

[10] S. Barthelmé, N. Tremblay, A. Gaudillière, L. Avena, and P.-O. Amblard. "Estimating the inverse trace using random forests on graphs". In: *XXVIIème colloque GRETSI*. 2019. eprint: 1905.02086.

[11] R. Bauerschmidt, N. Crawford, T. Helmuth, and A. Swan. "Random Spanning Forests and Hyperbolic Symmetry". In: *Communications in Mathematical Physics* 381.3 (2021), pp. 1223–1261.

[12] A. Bedini, S. Caracciolo, and A. Sportiello. "Phase transition in the spanning-hyperforest model on complete hypergraphs". In: *Nuclear Physics B* 822.3 (2009), pp. 493–516.

[13] I. Benjamini, H. Kesten, Y. Peres, and O. Schramm. "Geometry of the uniform spanning forest: Transitions in dimensions 4, 8, 12". In: *Annals of Mathematics* 160.2 (2004), pp. 465–491.

[14] I. Benjamini, R. Lyons, Y. Peres, and O. Schramm. "Uniform Spanning Forests". In: *Annals of Probability* 29.1 (2001), pp. 1–65.

[15] D. F. de Bernardini and S. Popov. "Russo's Formula for Random Interlacements". In: *Journal of Statistical Physics* 160.2 (2015), pp. 321–335.

[16] G. Birkhoff. "Tres observaciones sobre el algebra lineal". In: *Univ. Nac. Tucumán Rev. Ser. A* 5 (1946), pp. 147–151.

[17] R. Burton and R. Pemantle. "Local Characteristics, Entropy and Limit Theorems for Spanning Trees and Domino Tilings Via Transfer-Impedances". In: *Annals of Probability* 21.3 (1993), pp. 1329–1371.

[18] P. Chebotarev. "Spanning forests and the golden ratio". In: *Discrete Applied Mathematics* 156.5 (2008), pp. 813–821.

[19] P. Chebotarev and E. Shamis. "The Matrix-Forest Theorem and Measuring Relations in Small Social Groups". In: *Automation and Remote Control* 58.9 (1997), pp. 1505–1514.

[20] M. D'Achille, N. Enriquez, and P. Melotti. *Local limit of massive spanning forests on the complete graph*. 2024. arXiv: 2403.11740. URL: https://arxiv.org/abs/2403.11740.

[21] P. Diaconis and W. Fulton. *A growth model, a game, an algebra, Lagrange inversion, and characteristic classes*. Tech. rep. Stanford University, 1991.

[22] R. Dilworth. "A Decomposition Theorem for Partially Ordered Sets". In: *Ann. of Math.* 51.1 (1950), pp. 161–166.

[23] R. M. Dudley. "Distances of Probability Measures and Random Variables". In: *Ann. Math. Statist.* 39.5 (1968), pp. 1563–1572.

[24] R. M. Dudley. *Real Analysis and Probability*. 2nd ed. Cambridge University Press, 2002.

[25] T. Feder and M. Mihail. "Balanced matroids". In: *Proceedings of the twenty-fourth annual ACM symposium on theory of computing*. ACM, 1992, pp. 26–38.

[26] D. Feldman. "Doubly Stochastic Measures: Three Vignettes". In: *Distributions with Fixed Marginals and Related Topics*. 28. Institute of Mathematical Statistics, 1996, pp. 84–96.

[27] L. R Ford and D. R Fulkerson. "Network Flow and Systems of Representatives". In: *Canad. J. Math.* 10 (1958), pp. 78–84.

[28] L. R. Ford and D. R. Fulkerson. "Maximal flow through a network". In: *Canad. J. Math.* 8.3 (1956), pp. 399–404.

[29] G. Grimmett. *Percolation*. Berlin / Heidelberg: Springer, 1999.

[30] G. Grimmett. *The Random-Cluster Model*. Berlin / Heidelberg: Springer, 2006.

[31] G. Grimmett and S. N. Winkler. "Negative association in uniform forests and connected graphs". In: *Random Structures & Algorithms* 24.4 (2004), pp. 444–460.

[32] H. Guo, M. Jerrum, and J. Liu. "Uniform Sampling Through the Lovász Local Lemma". In: *Journal of the ACM* 66.3 (2019), pp. 1–31.

[33] P. Hall. “On representatives of subsets”. In: *J. Lond. Math. Soc.* 10.1 (1935), pp. 26–30.

[34] P. R. Halmos and H. E. Vaughan. “The Marriage Problem”. In: *Am. J. Math.* 72.1 (1950), pp. 214–215.

[35] J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. “Determinantal Processes and Independence”. In: *Probability surveys* 3 (2006). ISSN: 1549-5787.

[36] J. Hsu. “Probabilistic Couplings for Probabilistic Reasoning”. PhD thesis. 2017. arXiv: 1710.09951.

[37] T. Hutchcroft. “Interlacements and the wired uniform spanning forest”. In: *Annals of Probability* 46.2 (2018), p. 1170.

[38] T. Hutchcroft and A. Nachmias. “Indistinguishability of trees in uniform spanning forests”. In: *Probability Theory and Related Fields* 168.1-2 (2017), pp. 113–152.

[39] T. Hutchcroft and A. Nachmias. “Uniform Spanning Forests of Planar Graphs”. In: *Forum of Mathematics, Sigma* 7 (2019).

[40] B. D. Jones, B. G. Pittel, and J. S. Verducci. “Tree and Forest Weights and Their Application to Nonuniform Random Graphs”. In: *Annals of Applied Probability* 9.1 (1999), pp. 197–215.

[41] A. A. Járai, F. Redig, and E. Saada. “Approaching Criticality via the Zero Dissipation Limit in the Abelian Avalanche Model”. In: *Journal of Statistical Physics* 159.6 (2015), pp. 1369–1407.

[42] J. Kahn and M. Neiman. “Negative correlation and log-concavity”. In: *Random Structures & Algorithms* 37.3 (2010), pp. 367–388.

[43] R. Kenyon. “Lectures on dimers”. In: *Statistical Mechanics*. American Mathematical Society, 2009, pp. 191–230.

[44] R. Kenyon. “Spanning forests and the vector bundle Laplacian”. In: *Annals of Probability* 39.5 (2011), pp. 1983–2017.

[45] R. Kenyon. “Determinantal spanning forests on planar graphs”. In: *Annals of Probability* 47.2 (2019).

[46] J. Kingman. “Completely random measures”. In: *Pacific journal of mathematics* 21.1 (1967), pp. 59–78.

[47] G. Kirchhoff. “Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird”. In: *Annalen der Physik* 148.12 (1847), pp. 497–508.

[48] V. Klee and C. Witzgall. “Facets and vertices of transporation polyhedra”. In: *Mathematics of the decision sciences, Part 1*. 1968, pp. 257–282.

[49] D. König. “Über Graphen und ihre Anwendung auf Determinantentheorie und Mengenlehre”. In: *Math. Ann.* 77.4 (1916), pp. 453–465.

[50] D. König. “Graphen und Matrizen”. In: *Mat. Fiz. Lapok* 38 (1931), pp. 116–119.

[51] T. Koperberg. “Couplings and Matchings: Combinatorial notes on Strassen’s theorem”. In: *Statistics & Probability Letters* 209 (2024), p. 110089.

[52] V. T. Koperberg. “Loop-erased partitioning of sparse graphs”. Leiden University, Master’s thesis. 2020.

[53] G. F. Lawler. “A self-avoiding random walk”. In: *Duke Math. J.* 47.3 (1980), pp. 655–693.

[54] G. Lawler, X. Sun, and W. Wu. “Four-Dimensional Loop-Erased Random Walk”. In: *The Annals of probability* 47.6 (2019), pp. 3866–3910.

[55] Y. Le Jan. “Markov Loops and Renormalizaion”. In: *The Annals of probability* 38.3 (2010), pp. 1280–1319.

[56] Y. Le Jan. *Markov Paths, Loops and Fields: École d’Été de Probabilités de Saint-Flour XXXVIII - 2008*. Springer Nature, 2011.

[57] T. Lindvall. “On Strassen’s theorem on stochastic domination”. In: *Electron. Commun. Prob.* 4 (1999), pp. 51–59.

[58] L. Lovász and M. D. Plummer. *Matching Theory*. North-Holland Mathematics Studies. Elsevier Science, 1986.

[59] T. Lupu. “From Loop Clusters and Random Interlacements to the Free Field”. In: *The Annals of Probability* 44.3 (2016), pp. 2117–2146.

[60] R. Lyons and Y. Peres. *Probability on Trees and Networks*. Cambridge University Press, 2016.

[61] P. Marchal. “Loop-Erased Random Walks, Spanning Trees and Hamiltonian Cycles”. In: *Electronic Communications in Probability* 5 (2000), pp. 39–50.

[62] J. McKee and C. Smyth. *Around the unit circle*. Springer Nature Switzerland AG, 2021.

[63] K. Menger. “Zur allgemeinen Kurventheorie”. In: *Fund. Math.* 10.1 (1927), pp. 96–115.

[64] P. van Mieghem. *Graph Spectra for Complex Networks*. Cambridge University Press, 2010.

[65] L. Mirsky. “Hall’s criterion as a ‘self-refining’ result”. In: *Monatshefte für Mathematik* 73.2 (1969), pp. 139–146.

[66] O. Ore. “Graphs and matching theorems”. In: *Duke Math. J.* 22.4 (1955), pp. 625–639.

[67] R. Pemantle. “Choosing a Spanning Tree for the Integer Lattice Uniformly”. In: *Annals of Probability* 19.4 (1991), pp. 1559–1574.

[68] R. Pemantle. “Towards a theory of negative dependence”. In: *Journal of Mathematical Physics* 41.3 (2000), pp. 1371–1390.

[69] Y. Y. Pilavci, P.-O. Amblard, S. Barthelmé, and N. Tremblay. “Smoothing graph signals via random spanning forests”. In: *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 5630–5634.

[70] Y. Y. Pilavci, P.-O. Amblard, S. Barthelmé, and N. Tremblay. “Graph Tikhonov Regularization and Interpolation via Random Spanning Forests”. In: *IEEE Transactions on Signal and Information Processing over Networks* 7 (2021), pp. 359–374.

[71] J. Pitman. "Coalescent Random Forests". In: *Journal of Combinatorial Theory. Series A* 85.2 (1999), pp. 165–193.

[72] J. Pitman. *Combinatorial Stochastic Processes*. Springer, 2006.

[73] P. F. Reichmeider. *The equivalence of some combinatorial matching theorems*. Polygonal Pub. House, 1984.

[74] O. Schramm. "Scaling limits of loop-erased random walks and uniform spanning trees". In: *Israel Journal of Mathematics* 118.1 (2000), pp. 221–288.

[75] V. Strassen. "The existence of probability measures with given marginals". In: *Ann. Math. Stat.* 36 (1965), pp. 423–439.

[76] W. Tutte. "Graph-polynomials". In: *Advances in Applied Mathematics* 32.1 (2004), pp. 5–9.

[77] D. B. Wilson. "Generating random spanning trees more quickly than the cover time". In: *Proceedings of the Twenty-Eight Annual ACM Symposium on the Theory of Computing*. Vol. 96. 1996, pp. 296–303.

I must acknowledge myself to be indeed a very backward scholar; since I cannot now discover an argument which, it seems, was perfectly familiar to me long before I was out of my cradle.

- Hume; An Enquiry Concerning Human Understanding

Samenvatting

Grafen zijn wiskundige representaties van netwerken. Denk als voorbeelden van netwerken aan het wegennet bestaande uit splitsingen en kruisingen die met wegen zijn verbonden, sociale netwerken van mensen verbonden door vrienschapsrelaties of aan het netwerk van neuronverbindingen dat zich in de hersenen bevindt. Grafen modelleren deze netwerken als een verzameling punten, *knopen* genaamd, met een bijbehorende verzameling *lijnen*, waarvan elke lijn twee knopen met elkaar verbindt. De lijnen duiden de verbindingen aan die in het netwerk voorkomen. Het analyseren van de wiskundige eigenschappen van grafen kan inzichten verschaffen in vraagstukken waarin een netwerkstructuur voorkomt. Voorbeelden van dergelijke vraagstukken zijn of het sluiten van een specifieke weg verkeersopstoppingen zal veroorzaken of juist zal verhelpen, of hoe snel een gerucht of virus zich door een groep mensen zal verspreiden. Echter, voor grafen met een groot aantal knopen en lijnen kan het oplossen van vraagstukken dusdanig veel berekeningen vergen, dat dit zelfs met behulp van moderne computers niet haalbaar is.

Bossen zijn een speciaal type graaf met een simpele structuur: voor alle mogelijke tweetallen knopen bestaat er hoogstens één pad, een aaneenschakeling van lijnen, tussen deze twee knopen, oftewel een bos bevat geen *kringen*. De knopen van een bosgraaf kunnen worden onderverdeeld in één of meerdere groepen genaamd *samenhangscomponenten*, waarbij knopen tot verschillende componenten behoren wanneer ze niet verbonden zijn via een pad van lijnen. Een samenhangscomponent van een bos wordt ook wel een *boom* genoemd. De terminologie 'bossen' en 'bomen' is afgeleid van de grafische representatie van dergelijke grafen. Een afbeelding van een bosgraaf lijkt enigsinds op een verzameling bomen met een stam en met takken. Wegens hun relatief simpele structuur zijn bossen gemakkelijker te analyseren dan gecompliceerdere grafen. Een complexe graaf bevat meerdere deelgrafen met een bosstructuur. Deze deelgrafen kunnen worden verkregen uit de oorspronkelijke graaf door alle knopen te behouden en sommige van de lijnen te verwijderen. Een dergelijke bosdeelgraaf bevat informatie over de oorspronkelijke graaf, en zou hierdoor kunnen helpen bij het ofwel exact berekenen ofwel benaderen van de relevante eigenschappen van de graaf. Echter, in veel gevallen gaat in de reductie van complexe graaf tot bosgraaf te veel informatie verloren. Een mogelijke oplossing is om random bosdeelgrafen te gebruiken. De complexiteit van de graaf kan dan deels worden gevat in de kansverdeling van het random bos, en niet alleen in de structuur van het bos zelf.

Het bovengenoemde idee om ingewikkelde grafen te benaderen met random bosdeelgrafen vormt een belangrijke motivatie voor het onderzoek in dit werk. In het eerste deel van dit proefschrift bestuderen we een specifieke kansverdeling op bosdeelgrafen, de zogenaamde *Kirchhoff bosgraafkansverdeling*. Deze kansverdeling leent zich bij uitstek voor mogelijke toepassingen in de analyse van complexe netwerken, door

het bestaan van een efficiënte methode, genaamd *Wilsons algoritme*, waarmee een steekproef met deze kansverdeling genomen kan worden. In dit huidige werk worden Kirchhoff bosgrafen niet direct gebruikt voor het analyseren van complexe netwerken. In plaats daarvan wordt gepoogd een beter theoretisch begrip te krijgen van Kirchhoff bosgrafen, waardoor mogelijk toegepaste technieken in de netwerkanalyse zouden kunnen worden ontwikkeld op basis van dit fundamentele onderzoek.

Wilsons algoritme kan niet alleen worden gebruikt als methode om efficiënt steekproeven mee te nemen. Het algoritme gebruikt *random walks* op de oorspronkelijke graaf om een random bosgraaf volgens de Kirchhoff verdeling te construeren. De kringen die in het traject van een random walk voorkomen worden verwijderd, zodat de overgebleven doorlopen lijnen een bos vormen. Hierdoor kan Wilsons procedure ook gebruikt worden als theoretisch hulpmiddel om vragen over Kirchhoff bosgrafen te vertalen naar vragen over random walks, die eenvoudiger te bestuderen zijn.

De Kirchhoff bosgraafverdeling hangt af van een *intensiteitsparameter*, die kan worden aangepast om het verwachte aantal lijnen in een Kirchhoff bos te bepalen. In hoofdstuk 2 bestuderen we de kans dat twee knopen onderdeel zijn van dezelfde boom in een Kirchhoff bos, en in het bijzonder hoe deze kans afhangt van de intensiteitsparameter.

Hoofdstukken 3 en 4 richten zich op een ander aspect van Wilsons algoritme dan het resulterende bos. In hoofdstuk 3 ligt de nadruk op het *bezettingsveld* van Wilsons procedure, dat voor elke knoop telt hoe vaak deze in totaal door alle random walks is bezocht. Dit bezettingsveld is nauw verwant aan het bezettingsveld van een *random walk loop-soup*, en aan andere modellen in de statistische fysica zoals de *discrete Gaussian free field*. Door niet alleen te kijken naar het bezettingsveld maar naar de verzameling verwijderde kringen, worden in hoofdstuk 4 worden de resultaten uit hoofdstuk 3 uitgebreid. Hier laten we zien dat door Kirchhoff bosverdelingen met verschillende intensiteiten te *koppelen*, deze gekoppelde variant van Wilsons algoritme kan worden gebruikt om een loop-soup te construeren.

Twee kansverdelingen zijn niet noodzakelijkerwijs *onafhankelijk* van elkaar. Als los van elkaar een zeszijdige dobbelsteen en een twaalfzijdige dobbelsteen worden gegooid, dan zijn beide uitkomsten onafhankelijk. Dat wil zeggen, het resultaat van de ene worp geeft geen informatie over het resultaat van de andere. Het is echter mogelijk om de kansverdeling van een zeszijdige dobbelsteen na te bootsen door de uitkomst van de twaalfzijdige dobbelsteen te halveren en naar boven af te ronden. De verdeling van de nagebootste zeszijdige dobbelsteen is dan precies hetzelfde als die van een fysieke zeszijdige dobbelsteen, maar is niet langer onafhankelijk van de uitkomst van de twaalfzijdige. Dit is een voorbeeld van een *koppeling* van twee kansverdelingen, een algemene methode in de kansrekening om meerdere verdelingen gezamenlijk te beschouwen.

De *stelling van Strassen* is een vermaard kanstheoretisch resultaat dat aangeeft of twee verdelingen dusdanig te koppelen zijn dat deze koppeling een aantal gewenste eigenschappen heeft. In hoofdstuk chapter 5 wordt een nieuw eenvoudig bewijs van deze stelling gegeven voor het specifieke geval dat beide verdelingen een eindig aantal uitkomsten hebben. Dit bewijs vertaalt het koppelingsprobleem naar een grafentheo-

retisch probleem, en laat zien dat alle relevante informatie van de resulterende graaf gevat kan worden in één enkele bosdeelgraaf. Verder belicht dit bewijs de connectie tussen Strassens stelling en de *huwelijksstelling van Hall*, een bekend resultaat dat betrekking heeft op *matchings* in bipartiete grafen.

Summary

Graphs are mathematical abstractions that represent network structures. As examples of network structures you can think of road networks consisting of junctions and crossings connected by roads, social networks of people connected by friendships or the network of interconnected neurons in your brain. Graphs model networks as a set of points and an associated set of lines, with each line connecting two points. These lines represent the connections that occur in the network. For historical reasons, the points of a graph are called *vertices*, and the lines are called *edges*. Analyzing the mathematical properties of graphs gives insight into problems that involve networks. Examples of such problems are whether closing a specific road will increase or decrease traffic congestion, or how fast a rumor or virus will spread among a certain group of people. However, for graphs containing a large number of vertices and edges, finding solutions to such problems requires an amount of computation that, even with the help of modern computers, is unfeasible.

Forests are a special type of graph with a simple structure: between any pair of vertices there exists at most one path traversing the edges of the graph, i.e., these graphs contain no *cycles*. The vertices of a forest can be divided into one or more groups, called *connected components*, where vertices belong to different components if they are not connected via a path of edges. One such component is also called a *tree*. The terms 'forest' and 'tree' are derived from their graphical representation, since a drawing of a forest graph somewhat resembles a collection of trees with a trunk and with branches. Due to their relative simplicity, forests can be analyzed more efficiently than more complicated graphs. A complex graph contains many forest subgraphs, which can be obtained from the original graph by keeping all of its vertices and removing some of its edges. Such a forest subgraph contains information on the original graph, and hence could be of help in computing, or at least approximating, properties of interest. In many cases, however, too much information of the graph is lost in the reduction of the graph to a forest graph. A possible remedy is to use random forests, sampled from the full set of forest subgraphs, to approximate properties of interest. In this way, part of the complexity of the graph can be captured by the randomness of the forest rather than the structure of the sampled forests itself.

The above idea of using random forests to approximate complicated graphs provides one of the main motivations for the present thesis. In the first part of the thesis we study a specific probability distribution on forest subgraphs, which recently has been called the *Kirchhoff forest distribution*. This distribution lends itself particularly well for possible applications in network analysis, due to an efficient sampling procedure, called *Wilson's algorithm*, that can be used to sample forests from this distribution. Rather than applying Kirchhoff forests for network analysis, this thesis aims at providing a better theoretical understanding of the Kirchhoff forest distribu-

tion, thus laying a foundation on which applied techniques in network analysis can be built.

Wilson's algorithm is not only useful as an efficient sampling procedure. The algorithm constructs a random forest according to the Kirchhoff forest distribution by using *random walks* on the original graph. The cycles that appear in the edges that are traversed by a random walk are removed, so that the remaining edges form a forest. Therefore, Wilson's procedure can be used as a theoretical tool to translate questions about Kirchhoff forests into questions about random walks, which are simpler to study.

The Kirchhoff forest distribution depends on an *intensity parameter*, which can be tuned to fix the expected number of edges in a Kirchhoff forest. In chapter 2 we study the probability that two vertices are part of the same tree of a Kirchhoff forest, in particular, how this probability depends on the intensity parameter.

Chapters 3 and 4 focus on a different aspect of Wilson's algorithm. In chapter 3 the emphasis lies on the *occupation field* of Wilson's sampling procedure, which for each vertex counts the total number of visits by all random walks that are used in the construction of a Kirchhoff forest. This occupation field is closely related to the occupation field of a random walk *loop-soup*, and to other well-known models in statistical physics, such as the *discrete Gaussian free field*. In chapter 4 we extend the results obtained in chapter 3 by considering not only the occupation field, but the set of removed cycles. We show that, by coupling together Kirchhoff forests of different intensities, the coupled version of Wilson's algorithm can be used to construct a loop-soup.

Two probability distributions need not be *independent* from each other. If you throw both a six-sided die and a twelve sided die, then the resulting outcomes are independent, i.e., the result of one die does not provide any information on the result of the other. However, it is possible to emulate the distribution of the six-sided die by halving the result from a twelve-sided die, rounding up the outcome to an integer. The distribution of the emulated six-sided die is then exactly the same as that of a six-sided die, but it is no longer independent of the twelve-sided die. This is an example of a *coupling* of two distributions, a general method in probability theory to observe multiple distributions together.

Strassen's theorem is a celebrated result that tells you whether it is possible to construct a coupling of two distributions with certain desired properties. In chapter 5 a novel elementary proof of this theorem is provided for the special case of two distributions with finitely many outcomes. This proof translates the coupling problem into a graph problem, and shows that all relevant information of the resulting graph can be captured by a single forest subgraph. The proof highlights the connection between Strassen's theorem and *Hall's marriage theorem*, a well-known result concerning *matchings* in bipartite graphs.

Acknowledgements

This thesis could not have been realized without the help of the many people who have supported me along the way. Here I would like to thank some of them, although I will not be able to give sufficient thanks to everyone involved.

Let me start by expressing mayor gratitude towards my supervisors **Luca Avena**, **Alexandre Gaudillière** and **Frank den Hollander**.

In particular, I am grateful to Luca not only for supervising my PhD but for guiding my entire mathematical career over the past decade. He introduced me to mathematical research, guided me through bureaucratic difficulties, and made me grow not only as a mathematician but as a person. I'll be forever grateful for meeting him, as without him my life would not have been the same.

I thank Alex for our many brainstorming sessions before the blackboard in which our communication seemed effortless, and in which he seemed to grasp my vague ideas even before I did so myself. I also thank him for his warm welcome to Marseille during my internship. I still feel at home there whenever I visit. The writing of this thesis has not been without obstacles. It has been greatly facilitated by Alex' warm and dedicated support, for which I am truly grateful.

I'd like to thank Frank for taking on the responsibility of being my promotor and for the extensive feedback he provided on this manuscript.

The support of Alex, Luca and Frank has been invaluable, and their patience with me showed no bounds. I could not have wished for better supervisors.

Special acknowledgements go out to the members of the reading committee **Nathanaël Enriquez**, **Wioletta Ruszel** and **Christophe Sabot** for their time to read this thesis.

I further like to thank:

- the **NETWORKS** program for funding my research and its organizers for creating a stimulating and collaborative atmosphere during the training weeks;
- my close colleagues **Oliver Nagy**, **Daoyi Wang**, **Federico Capannoli**, **Pierfrancesco Dionigi**, and **Nandan Malhotra** for our pleasant meetings and inspiring discussions, and for making the training weeks truly enjoyable;
- **Ad de Stoppelaar** for his efforts to keep me motivated and oriented towards finishing this thesis;
- my father **Ilbert**, mother **Mary** and sister **Majsa** for providing me with the safety, security and support that allowed me to set out on this journey at all;
- **Theresa Song Loong** for her unconditional love, her understanding, and for always keeping me focussed on the important things in life, even during its difficult periods.

Curriculum Vitae

Twan Koperberg was born in Amsterdam in 1989. He completed his pre-university education at the Stedelijk Gymnasium in Leiden.

He obtained his bachelor's degree in Mathematics from Leiden University in 2017. In 2020 he received his master's degree in Mathematics from the same university, where he wrote his master's thesis titled 'Loop-erased partitioning of sparse graphs' under supervision of Luca Avena.

For his PhD Twan joined the probability group of the Mathematical Institute of Leiden University under the supervision of Frank den Hollander. There, he continued working on the research project initiated during his master's in a joint collaboration with Luca Avena and Alexandre Gaudillière, which culminated in the present thesis. His research was part of the NETWORKS consortium, a national research program on complex networks funded by the Ministry of Education, Culture and Science.

During his PhD Twan did a three month internship at Université d'Aix-Marseille. In addition to his research, Twan served as a teaching assistant for various courses taught at Leiden University.