Probabilistic graph inspections through forests

Koperberg, V.T.

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.