Probabilistic graph inspections through forests

Koperberg, V.T.

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.