Proposed Research – Fragmentation Trees: Douglas Rizzolo

Key Words: Fragmentation Trees, Continuum Random Trees, Phylogenetics, Cladograms

Much of my research as an undergraduate dealt with graphs of one variety or another. The constructive fixed point search algorithm I developed with Professor Su is based on tracing paths through particular graphs. In a separate project we also studied topological properties of metric trees and cycles. My work with Professor Strichartz at Cornell also focused on graphs; in particular on the relationship between finite graphs and the fractal structures that they can be used to approximate. While working on these disparate topics, whose common thread is using finite graphs study the infinite objects they approximate, I began think about the question: What kinds of objects arise as limits of graphs? U.C. Berkeley was the perfect place to ask this question because numerous members of the faculty work on various aspects of these limits. Shortly after arriving at U.C. Berkeley I began working with Professor James Pitman, who has pioneered the study of limits of fragmentation trees.

A tree $t_n$ with $n$ leaves is called a fragmentation tree on $n$ leaves if $t_n$ satisfies the following properties: (i) one internal (non-leaf) vertex is labelled as the root, (ii) except for possibly the root, $t_n$ has no degree 2 vertices, (iii) the $n$ leaves are labeled by $[n] := \{1, \ldots, n\}$, and (iv) each internal vertex is labeled by the set of leaves above it. The set of vertices of $t_n$ connected to the root then corresponds to a partition of $[n]$ via the labels of these vertices and, in this fashion, the entire tree corresponds a way of recursively partitioning, i.e. fragmenting, $[n]$. A fragmentation model is the assignment, for each $n$, of a probability distribution to the set of fragmentation trees with $n$ leaves. Fixing a fragmentation model, a fragmentation process is a sequence $(T_n)_{n=1}^\infty$ of trees where $T_n$ is a fragmentation tree with $n$ leaves chosen randomly according to the probability given by the model. Of particular interest are Markovian fragmentation models, where the probabilities come from a splitting kernel that for each vertex gives you the probability that the labels of the adjacent vertices above it have particular cardinalities. Such trees arise in biology as phylogenetic trees and their properties are used to gather information about the evolutionary process (see e.g. [2] for a survey). Fragmentation trees also occur in fields such as statistical classification theory, where they can be used to determine whether or not a classification structure contains information [3].

Natural probabilistic questions arise about the asymptotics of Markovian fragmentation processes. Of particular interest is whether or not these processes converge to, or embed in, some limiting tree. This notion can made rigorous by assigning trees edge lengths and using the continuum random trees introduced by Professor David Aldous in [1], which can be thought of as fractal tree structures with a continuum of leaves. Very recently, in [4], it was shown that, under a strong sampling consistency assumption and appropriate scaling, fairly general Markovian fragmentation processes converge in probability to a continuum random tree with respect to the Gromov-Hausdorff metric. I propose to study what happens when the strong sampling consistency hypothesis is removed.

This case is of great interest because many combinatorially motivated fragmentation models are not sampling consistent. For example, the model where for each $n$ the probability on fragmentation trees with $n$ leaves is uniform can be shown to be a Markov fragmentation model that is not consistent. Nonetheless, because of parallels with Aldous’s theory of asymptotics of another class of random trees called simply generated trees, it is conjectured that this uniform model converges in distribution (but not necessarily in probability) to a continuum random tree limit. I intend to prove this conjecture.
It is shown in [4] that the question of convergence to continuum limit is closely tied to the question: If a tree is chosen randomly according to the fragmentation model and then a leaf is chosen uniformly at random from the leaves of this tree, how high is it? i.e., how many edges are on the path from the root to this leaf? Over the past year, I have used methods from complex analysis to show that when scaled by $n^{-1/2}$ the distribution of heights of leaves converges to a Rayleigh distribution. In fact, I have shown this for a broad class of combinatorially motivated models that includes the uniform model as a special case. A natural next question is: Do the scaled moments converge as well? Proving that the answer to this question is affirmative seems well within reach and completing this proof will open the door to using the tools developed in [4] to prove convergence to a continuum limit. Thus I propose to first prove convergence of the scaled moments and then to try to leverage this result into a proof of convergence of these fragmentation models to a continuum limit.

The next step is then to use the insights into the inconsistent case provided by the uniform model to study the convergence Markovian fragmentations in general. The ultimate goal of this program of research is a complete characterization of the convergence of Markovian fragmentation trees to continuum limits, something akin to the known theory of convergence in distribution of partial sums of independent random variables. While the expected limits have been identified, much remains to be done in terms of establishing conditions for convergence. At present, I do not have all the tools necessary to work at this level of generality and I propose to acquire them over the coming months through two methods: (i) Taking directed reading courses with Professor Pitman that are focused on the current literature, and (ii) attending probability summer schools centered on this field.

Thus I intend work on several questions that I currently have the tools to answer while simultaneously acquiring the background needed address broader issues in the field. Furthermore, given that this field of research is quite young, a key aspect of making progress will be getting more people involved in the field. This will be accomplished through several means. First and foremost, all results will be submitted to peer reviewed journals as well as posted on arXiv. Having open access to them on arXiv will maximize the dissemination of the results, especially among groups that do not have access to peer reviewed journals that often come with a hefty price tag. In addition to regular posting of results, I also intend to give talks about this at relevant research conferences in order to bring the field to the attention of researchers working on related topics.

References


