Community detection via random spanning forests

Project co-supervised by Dr. L. Avena (MI) and Dr. D. Garlaschelli (LION)

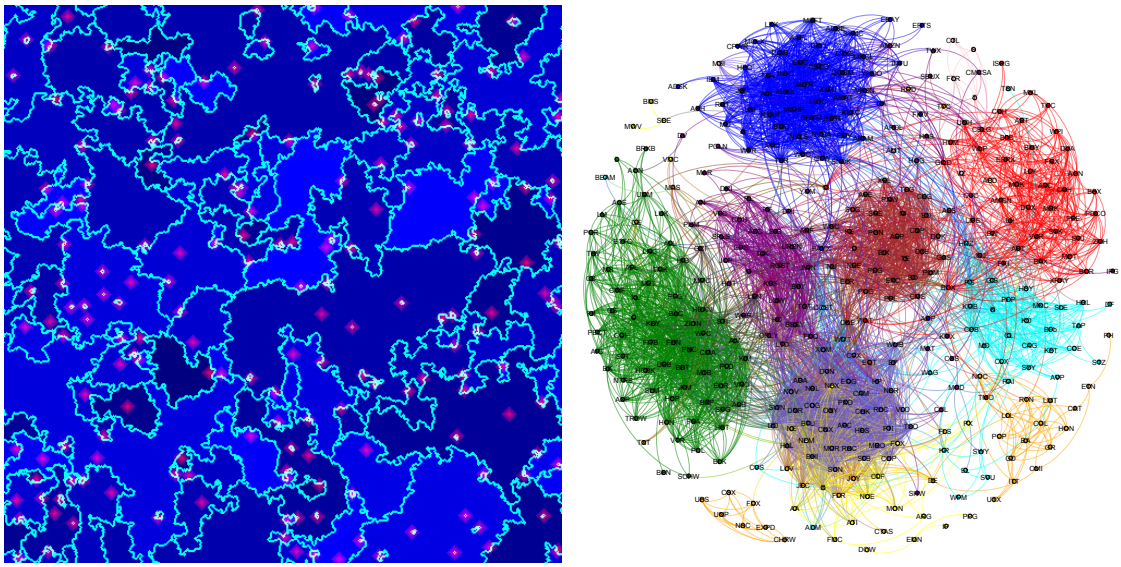
The goal of this project is to connect two currently unrelated lines of research: on the mathematical side, the theory of random spanning forests; on the physical side, the notion of community detection and its application to time series clustering.

In graph theory, a forest is a set of trees, each of which is defined as a connected graph with no loops. In a recent publication [1], the authors introduce and study a certain probability measure $\mathbb{P}$ on the set of spanning rooted oriented forests of a given finite connected weighted oriented graph. The random forest Φ sampled from the measure $\mathbb{P}$ induces a "good" random partition of the given graph. Figure 1 (left) illustrates a partition induced by a realization of this random forest when the graph is a square lattice box with unitary weights. The set of roots $\rho(\Phi)$, i.e., the set of points which are the roots of the trees constituting Φ (the distinguished red points in Fig. 1), has been shown to be a determinantal process. Intuitively, this means that these roots are points that repel each other (they are negatively correlated) and therefore tend to be "well" distributed in the graph. For example, by considering the random walk X associated to the weighted graph, it turns out that on average, no matter where X starts, it hits the set of roots $\rho(\Phi)$ in the same amount of time independently of the geometry of the graph. For practical purposes, an adaptation of the so called Wilson's algorithm provides an efficient procedure to sample realizations of these forests.

On the other hand, in the statistical physics community there is increasing interest towards the problem of *community detection*, i.e. the identification, in empirical networks, of sets of nodes that are more densely connected internally than with the rest of the network. A similar problem arises in the analysis of correlation matrices built from (e.g. financial or neural) time series: in this case, one wants to identify communities of time series that are more strongly correlated internally than with the other time series (see Fig. 1 right). Recently, it has been pointed out that the community detection methods that have been extensively developed for the analysis of networks are unsuitable for the analysis of correlation matrices, due to the different mathematical properties of these structures. This motivated the introduction of a novel community detection method that is suitable for the analysis of correlation matrices [2]. However, additional and alternative methods are still needed.

The aim of this project is that of proposing a novel correlation-based community detection method using the partition induced by the random spanning forests introduced in [1]. Empirical correlation matrices will be mapped to weighted graphs and the spanning forests of the latter will be interpreted as novel partitions into communities [2]. The project requires mathematical, computational and data-analysis work.

Figure 1: **Left:** a realization of a partition induced by a forest sampled from $\mathbb{P}$ on the two-dimensional 512×512 torus with unitary weights, using the method in [1]. Regions delimited by cyan lines correspond to the components identified by the trees in the forest. The roots of the trees in the forest are marked by red diamonds. **Right:** network obtained by filtering an empirical correlation matrix of financial time series. The original data are the time series of the stocks of the S&P 500 financial index, and the colors refer to the industrial sector to which the stocks belong. Communities are visible as subsets of nodes that are more densely connected (i.e. more strongly correlated) internally than with the rest of the system. These communities are however difficult to define mathematically and to find computationally. After [2].



References

[1] L. Avena and A. Gaudillière, On some random forests with determinantal roots, *ArXiv:1310.1723*, WIAS Preprint No. 1881, (2013).

[2] M. MacMahon and D. Garlaschelli, Community detection for correlation matrices, *ArXiv:1311.1924*, accepted for publication on *Phys. Rev. X* (2015).