Objectives of the internship

This report succintly presents the work done during the internship concluding the Master 2 program MVA - *Mathématiques, Vision, Apprentissage*. The internship was academic in nature, and lasted six months. It will be followed by a PhD thesis on the same subject, by the same intern, under the same advisors. Its goals were, sequentially,

1st To introduce the intern to the problem of community detection on graphs. This consists on getting a firm understanding of the different canonical approaches to the problem, gaining familiarity first with the classical and then with the recent research literature of the field, and gaining probabilistic as well as statistical intuitions that might be generalized to related problems;

2nd To approach a current research question, suggested by the advisors of the internship;

3rd To apply the knowledge obtained to think of new research directions autonomously, preparing the start of the intern's PhD.

Motivations for the internship

The field of community detection on graphs is rich both in theory and in applications. In this section, the motivations for the field and for the particular problem of the internship are explained.

Random graphs with communities

A graph is a mathematical object expressing the interaction between entities. These interactions might be deterministic, as in the example of a network of proteins, or they might be random, as when people of the same or different neighborhoods meet.

A random graph is a graph arising as a sample of a probabilistic model on the set of possible edges and possible attributes on the vertices of a graph. One classical example is the Erdős-Rényi model, denoted $G(n, p)$, over graphs with n vertices. Under this model, the connection between each pair of nodes with an edge is determined by a Bernoulli random variable of parameter p. This means that for each pair of vertices one connects them with probability p, and with a remaining probability of $1 - p$ one does not connect them. This is arguably the simplest random graph model possible, and it barely possesses any *structure* at all.

The presence of *communities* on a graph representing individuals and their interactions is of the most interesting and relevant such *structures* that a network can have. However, defining what a community actually is is delicate and not agreed upon. One

common intuition is to think of it as a set of vertices having more connections between themselves than with all other vertices. Such an intuition is useful in many situations, and it is called an *assortative* notion of community. In other cases, however, one's intuition of what a community is does not fit the assortative case. Consider a party, where people dance in pairs. There are fifty men and fifty women, and assume that each man will pair up with some woman to dance. In this case, one can consider that there are two communities in the party, men and women. However, no two members of the same community connect. This is what is called an *dissortative* notion of community, as its members share a *pattern* of connection instead of denser connections. The conceptual difficulties do not stop here, as frequently the notion of what is a community depends on the *scale* (i.e., the amount of "zoom" into the graph) considered.

One way of dealing with these conceptual difficulties is to consider a model of random graphs having a definite, *ground truth* attribution of communities to its nodes. Then, developing and studying different approaches to community detection based on graphs coming from such a model, before applying to real work networks. A popular model for this is called the Stochastic Block Model (SBM), and there are many variants associated to it.

Having fixed the SBM as the base model to develop and compare community detection algorithms, in this internship two main approaches of developing such algorithms were considered.

Probabilistic approaches

In *probabilistic* approaches to community detection, one assumes the SBM model for the graph observed, and uses algorithms derived from analyzing its likelihood function. In fact, the likelihood of an observation under the SBM is intractable, since computing it requires one to sum over all possible cluster assignments for the nodes (which are *latent*, unobserved variables of the model). Therefore, a common way around this is to substitute the exact likelihood function by a variational approximation. This approximation can be built in various ways, the simplest one being known as the *mean-field* approximation.

Spectral approaches

In *spectral* approaches, one derives algorithms based on heuristics and approximations to graph cut problems, without assuming the graph observed comes from a SBM, nor making use of its likelihood function. The idea behind this is to use some *graph operator* (classicaly the graph laplacian) in order to build a vectorial representation of the graph. In it, each node of the graph correspond to a vector. If such a representation is good, then clustering its vectors would correspond to clustering the nodes of the graph. This

clustering would be done with some classical algorithm of euclidean clustering, such as k-means. These approaches are very popular, but as discussed previously, without a ground truth for the communities there is no way of measuring their accuracies. This is where the link with the SBM appears, as it can be used as a base model to understand *why* and *when* spectral approaches work, and when they do not, what to do to fix it.

The starting problem of this internship

When