With over 6000 tweets per second being posted on Twitter, Google processing over 40000 queries every second, and more than 400 hours worth of youtube videos uploaded every minute, there is an urgent need to develop dynamic algorithms to make sense of massive amounts of data. However, despite the significant efforts made by the research community, there are very few dynamic data mining and machine learning algorithms which are efficient in practice and come with strong theoretical guarantees.

This internship proposal focuses on clustering problems, which have received significant attention in recent years [LV17, GKL17, MBM]. In particular, we shall focus on studying the dynamic $k$-means [AV07] and $k$-center [COP03] problems both from a theoretical and practical point of view. There are several “static” algorithms for those problems with strong theoretical guarantees [AV07, COP03] and which are used in practice. A fully dynamic algorithm for $k$-center has been developed in [CGS17].

Given any static algorithm, one could obtain a dynamic algorithm by running the static one from scratch whenever a data point is added or removed. Of course, such an algorithm would not be efficient. Typically, one wishes to handle deletions and insertions while requiring only a small amount of operations per update operation on average (e.g. constant or logarithmic on the maximum size of the input), without affecting the theoretical guarantees on the quality of the solution “too much”.

We shall focus on two main models of computation: the sliding window model [ELS15] and the fully dynamic model [CGS17]. In the former one, at any point in time only the most recent data items are retained (e.g. the most recent $n$ data points), while in the latter one points can be added or removed, arbitrarily. We shall also consider the variant where at most $l$ points (so-called outliers) can be removed from the input, in order to improve the quality of the solution.

During the internship the student will focus on one or more of the following aspects:

- develop a dynamic algorithm for the $k$-means in the sliding window or fully dynamic model of computations. Develop a dynamic algorithm for $k$-centers with outliers or improve the results presented in [CGS17]. Ideally, the student will be able to prove strong theoretical guarantees on the quality of the solution as well as on the number of operations required for each update operation (on average).
- implement the algorithm proposed (in C or Java) and conduct an evaluation of such an algorithm on real-world data (from Twitter and other highly dynamic social media).
- study the problem of dynamic clustering on graphs.
**Required skills.** For this internship it is required a strong expertise on theory of algorithms or excellent coding skills in C or Java. Depending on the interest of the students the internship will focus more on theoretical or practical issues.

The topic of this internship is related to the course on graph mining offered at MPRI. Participation to such a course is encouraged but not strictly required. Other research problems in the context of dynamic massive data mining and machine learning algorithms are available upon requests. A successful internship on this topic can lead to a PhD thesis.

**References**


[^1](http://wikimpri.dptinfo.ens-cachan.fr/doku.php?id=cours:c-2-29-2)