**PhD title**: Graph Compression

**Host laboratory**: Laboratoire d’Informatique de Bourgogne (LIB) – EA 7534 and Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS, UMR 5205 CNRS)

**Speciality of PhD**: Computer Science

**Keywords**: Graph, Compression, Algorithm, Data Science

**Job description**: In our world of Big Data, we are faced with a major issue of processing huge graphs. One approach considered in the ANR project Coregraphie is to build a summary (or compressed version) of the graph and to query this summary instead of the original graph.

Two main compression methods can be distinguished: Lossless and lossy compression. Lossless compression (or compact coding) decreases the size of coding of the graph without losing any information while controlling the cost induced by this coding on the operations, like for instance with WebGraph [BV04]. Lossy compression allows a part of the information (nodes and/or edges) to be lost. If the compression is by deleting edges, with speak about sparsification. This subclass of lossy compression allows more liberty on requests, like estimating the distance between two nodes [KB+21].

Lossy compression is mainly accomplished using two methods: Selecting a sample, i.e., a sub-graph using different technics (random walks, propagation, filtering, etc.) [HL13] and grouping nodes/edges (generalization) [CR15]. In most cases, lossy compression methods are specialized for one type of request [FL+12].

One major issue for lossy compression is then to determine to which extent the compression algorithms damage the initial graph and how this damage can be measured and controlled.

This thesis aims to concentrate on lossy compression by studying the impact of compression methods on the properties of the graph and/or requests that need to be performed on it.

We propose approaching this problem starting with simple requests like testing neighborhood or proximity between nodes and then more complex ones like finding a clique of a given size, do a clustering [QK15], or partition the graph into independent sets [T19].

Among further aspects that can be explored are:

- The links with lossless compression and combined approaches;
- Links with structural properties, in particular with some orders/hierarchies ((k-shells, k-trusses, modular decomposition, twin-width, etc.);
- Links with communities and hubs [GC+20];
- Compromises between preserving properties and anonymizing [MRT20].

The main application domain will be social networks for which the LIB lab possesses the scientific environment to handle them. The links with some human science issues may also be considered.

The final goal is to propose, using a combination of compression methods, generic compression
schemes allowing to take into account many types of requests and to control the bias induced by the compression.

References:


Candidate Profile: Master degree or Ingeenier in computer science or mathematics, with strong skills in graph algorithms / network science / data mining and programming in Python/C++/Java.

Financing Institution: ANR Contract Coregraphie (ANR-20-CE23-0002)

Application deadline: June 25
Start of contract: October 2021

Supervisor(s): H. Chérifi (LIB, Dijon), H. Séba (LIRIS, Lyon), O. Togni (LIB, Dijon)
Contacts: hocine.cherifi@u-bourgogne.fr, hamida.seba@univ-lyon1.fr, olivier.togni@u-bourgogne.fr