****



|  |
| --- |
| **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  In this Big Data world, we face the central issue of processing massive graphs. The 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.  One can distinguish between lossless and lossy compression techniques. Lossless compression (or compact coding) decreases the size of the graph representation without losing any information while controlling the cost induced by coding on the operations, like, for instance, in WebGraph [BV04].  Lossy compression allows a part of the information (nodes and edges) to be lost. If the compression is by deleting edges, one speaks of sparsification. This subclass of lossy compression allows more effective requests, such as estimating the distance between two nodes [KB+21]. Lossy compression is mainly accomplished using two approaches: 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 determining 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. We plan to investigate the impact of compression methods on the graph topological properties and/or requests performed. We propose approaching this issue starting with simple requests such as testing neighborhoods or proximity between nodes. We will then study more complex requests such as finding a given size clique, clustering [QK15] or partitioning 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 community structure and centrality measures [GC+20];  • Compromises between preserving properties and anonymizing [MRT20].  The main application domain is social networks. Indeed, the LIB lab has the  scientific environment to handle massive online social data. Data from Medicine, Biology, Economics may also be considered. The final goal is to propose effective tailored compression methods and generic compression schemes to deal with many types of requests while controlling the bias induced by the compression. |
| **References :**  [BV04] P. Boldi and S. Vigna, The webgraph framework i: Compression techniques. In Proceedings of the 13th International Conference on World Wide Web, WWW ’04, pages 595–602, New York, NY, USA, 2004. ACM.  [CR15] J. Casa-Roma, F. Rousseau, Community-preserving generalization of social networks. In IEEE. 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). [S.l.], 2015. p. 1465–1472.  [FL+12] W. Fan, J. Li, X. Wang, and Y. Wu, Query preserving graph compression. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (SIGMOD '12), 2012. DOI:<https://doi.org/10.1145/2213836.2213855>  [GC+20] Z. Ghalmane,C. Cherifi, H. Cherifi, M. El Hassouni*.* Extracting backbones in weighted modular complex networks. *Sci Rep* **10,** 15539 (2020). <https://doi.org/10.1038/s41598-020-71876->  [HL13] P. Hu and WC. Lau, A survey and taxonomy of graph sampling. 2013. CoRR abs/1308.5865, <http://arxiv.org/abs/1308.5865,1308.5865>  [KB+21] A. E. Kiouche, J. Baste, M. Haddad and H.Seba. A Neighborhood-preserving Graph Summarization, 2021, coRR abs/ 2101.11559, https://arxiv.org/abs/2101.11559.    [MRT20] G. Minello, L. Rossi and A. Torsello, k-Anonymity on Graphs using the Szemerédi Regularity Lemma, IEEE Transactions on Network Science and Engineering, 2020, <https://doi.org/10.1109/TNSE.2020.3020329>.  [QK15] F. Queyroi and S. Kirgizov. Suppression distance computation for hierarchical clusterings. Information Processing Letters, Volume 115, Issue 9, 2015.  [T19] O. Togni. Coloring Large Real World Networks : the DSAT-ratio, In MARAMI 2019, 2019. [YAA21] M. I. Yousuf, I. Anwer and R.Anwar, Empirical Characterization of Graph Sampling Algorithms, 2021, coRR abs/2102.07980 , https://arxiv.org/abs/2102.07980 |
| **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](mailto:hocine.cherifi@u-bourgogne.fr), [hamida.seba@univ-lyon1.fr](mailto:hamida.seba@univ-lyon1.fr), [olivier.togni@u-bourgogne.fr](mailto:olivier.togni@u-bourgogne.fr) |