Sujet de thèse: *Structure and Algorithms for Community Detection Problems in Graphs*
Encadrants : Christophe Paul (DR1 CNRS),
Dimitrios THILIKOS TOULOUPAS (DR2 CNRS)
Equipe/Laboratoire : AlGCo/LIRMM (UMR 5506)

A *graph* (or network) is a simple mathematical concept expressing relationships between entities. Graphs constitute a cornerstone structure in many different domains. *Graph mining* is a subfield of data mining which deal with the efficient extraction and processing of knowledge from potentially large datasets that can be naturally encoded/modeled by graphs.

There are algorithmic tasks which typically arise in graph mining. A *community* is a set of nodes in a graph which is densely connected among its members while comparably less connected to nodes outside of it. Communities are a basic building block of a large-scale network and can be considered rather as an independent part of a network [7]. Communities emerge in almost all domains in which data sets are naturally expressed as graphs. Therefore, detecting a community, or *dense graph discovery*, is of great importance in Graph Mining. Viewing a community as a building block of a network naturally leads us to investigate the community structure. Understanding a large-scale network as a collection of (disjoint or overlapping) interlinked communities is the first step in understanding large-scale networks. The community structure as a collection of communities [7] is referred to as *clustering* in graph mining and *graph partitioning* in theoretical computer science. The concept of clustering has received a great deal of attention across disciplines such as computer vision, web graphs, and communication networks [9, 10].

Algorithmic and scalability. Typically, an *efficient algorithm* is an algorithm whose running time is a polynomial function of the size of the input. Most of computational problems (including many problems on graphs) are not expected to admit any efficient algorithm. This point of view comes in contrast to the fact that, in many practical cases (especially those dealing with huge data-sets organized by graphs) there still exist algorithmic approaches that can provide satisfactory solutions based on scalable implementations. For this reason, several alternative algorithmic paradigms have been proposed, based on more relaxed, definitions of algorithmic efficiency. The thesis will make use of the following two:

Parameterization. Parameterized algorithms are super-polynomial algorithms where the non-polynomial part of their running time exclusively depends only on some (typically small) key parameters of the problem. Such algorithms, when the parameter is small enough, can still be efficient in practice. Especially for problems on graphs, parameters reflect structural characteristics of the input graphs that are prevalent in real world applications. A typical example of such an algorithm, related to graph mining, was given in [4] for the Cluster Deletion problem asking whether it is possible to transform a m-edge graph to a collection of vertex-disjoint cliques by removing at most k edges. This problem can be solved in $O(1.62^k + m)$ steps [4]. Therefore, if the parameter k is small enough for the graphs that we are dealing with, we have an algorithm whose guarantied performance is practically linear in the magnitude of the graph. The potential of parameterized algorithm design techniques for the design of scalable algorithms on graphs representing massive data sets remains, to some extend, rather unexploited.

Preprocessing and data reduction. Here the target is to "reduce" the input of the problem as much as possible so that a brute force algorithm for it will have to deal with a considerably smaller/simpler instance. Preprocessing is present almost in every application of graph mining: it roughly consists in transforming the input graph in a much simpler one without altering significantly the quality of the knowledge it contains. Recently, a novel theoretical framework, called *kernelization*, was proposed as a solid mathematical formulation of preprocessing [8]. Kernelization concerns parameterized problems and asks for polynomial algorithms that can reduce instances of the problem to equivalent ones whose size depends *exclusively* on the parameter of the problem. For instance, it is known that the Cluster Deletion problem can be easily preprocessed in polynomial time so to produce an equivalent graph of $\leq 2k$ vertices [1]. It is a challenge to investigate up to which point such preprocessing algorithms can offer scalable algorithms for specific applications in graph mining.

Graph theory offers a wealth of structural results and a significant advance in algorithms design can be built on them. It has been long known that many problems admit much more efficient algorithms on several classes of sparse graphs. It appears that modern structural graph theory offer powerful algorithmic ideas which can be exploited for dealing with community detection and clustering problems in Graph Mining.

Objectives. The objective of the thesis is to make use and/or develop combinatorial tools dedicated to the design of graph clustering and community detection algorithms on large-scale networks. Some of the challenges that are expected to be met are the following:

A. Use alternatives of hierarchical decompositions. An example of structural hierarchization of a graph is the mention the notion of *k-core decomposition* [2] that partitions the graph into layers of

increasing density. It appears that the densest cores of a graph are roughly indicating its clustering structure and thus, many classical clustering algorithms can be significantly accelerated when they start from the densest core. That way, the k-core decomposition can be used as a "structural guide" for the implementation of any known clustering procedure. Communities can be seen as building blocks of a network, which subsume the idea that 'if a part of a network is dense, it must deliver important information that should not be cut short'. The thesis will study alternative hierarchizations using either variants of the k-core concept or other notions of "local density" in graphs. Also it will examine several other extensions of core structures combining different combination of degree and connectivity demands.

B. Alternative criteria for clustering. An interesting question on clustering is whether "density" is the best criterion. Actually, evidence suggests that this is not always the case as, sometimes, a cluster can be strongly interconnected even if it is relatively sparse. A typical graph-theoretical structure reflecting this behavior is then notion of expander graphs. The study of expander graphs is nowadays an active part of modern combinatorics with many algorithmic applications [5]. Currently, there is no developed graph theoretic concept for capturing the property of remaining "locally highy-connected" while maintaining low global (inter-) connectivity. Similar questions can be made on the interplay between local and global density in graphs. The proposed thesis will tackle such questions combining graph theoretic considerations with empirical/experimental knowledge from certain graph mining applications.

C. Clustering on more general types of graphs. Most of the approaches for studying clustering in graph mining concern simple graphs. However, in many cases, the semantics of data sets correspond to directed graphs or even more general structures such as hyper-graphs or colored graphs [3, 6]. Some of the questions that will be addressed are the following: What is a successful clustering algorithm on such structures? Is it possible to extend current parameterized/data reduction algorithms to this direction? Which graph theoretic concepts and parameters might be helpful for this? Which data sets can offer a good experimental base?

Research enviroment: The AlGCo team has a strong background on the proposed research program. AlGCo is one of the leading research teams in France in Parameterized Computation. Members of the team have contributed to international conferences and participated to research projects that are directly or indirectly related to parameterized complexity and have an increasing interest to kernelization and data reduction. (The publication record of the team for the last 5 years can be accessed at *http://www2.lirmm.fr/algco/publications.php*.

Research Calendar: During the first year, the research will be focused on the study/development of hierarchical decompositions of graphs for clustering applications. During the second year, alternative criteria for clustering will be examined. Finally, during the third year, the research will be directed to the study of clustering problems for digraphs and other more general combinatorial structures.

Candidate profile: The candidates should have a solid background in algorithms, complexity, and graph theory and strong motivation to work on related areas, including parameterized computation and approximation algorithms. Moreover, english language skills and knowledge/experience on programming in Python, C, and C++, will be appreciated.

References

[1] Y. Cao and J. Chen. *Parameterized and Exact Computation: 5th International Symposium, IPEC 2010*, chapter Cluster Editing: Kernelization Based on Edge Cuts, pages 60–71. Springer Berlin Heidelberg, 2010.

[2] C. Giatsidis, F. D. Malliaros, D. M. Thilikos, and M. Vazirgiannis. CoreCluster: A degeneracy based graph clustering framework. In *AAAI*, pages 44–50, 2014.

[3] C. Giatsidis, D. Thilikos, and M. Vazirgiannis. Evaluating cooperation in communities with the k-core structure. In *ASONAM*, pages 87–93, 2011.

[4] J. Gramm, J. Guo, F. Hüffner, and R. Niedermeier. Graph-modeled data clustering: Exact algorithms for clique generation. *Theory of Computing Systems*, 38(4):373–392, 2005.

[5] V. Karyotis, E. Stai, and S. Papavassiliou. *Evolutionary Dynamics of Complex Communications Networks*. CRC Press, Inc., Boca Raton, FL, USA, 2013.

[6] J. Kim and J.-G. Lee. Community detection in multi-layer graphs: A survey. *SIGMOD Rec.*, 44(3):37–48, Dec. 2015.

[7] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[8] D. Lokshtanov, N. Misra, and S. Saurabh. *The Multivariate Algorithmic Revolution & Beyond: Essays Dedicated to M. R. Fellows on the Occasion of His 60th Birthday*, chapter Kernelization – Preprocessing with a Guarantee, pages 129–161. Springer, 2012.

[9] M. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[10] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.