Università
Ca’Foscari
Venezia

Master’s Degree
in Computer Science

Final Thesis

Modularity Community Detection on the GPU

Supervisor

Ch. Prof. Claudio Lucchese

Graduand

Federico Fontolan

Matriculation Number

854230

Academic Year

2019 / 2020

Abstract

Modularity algorithms for the detection of communities are the de facto standard thanks to the fact that they offer the best result between efficiency and result. Moreover, these algorithms allow analyzing graphs much larger than those that can be analyzed with alternative techniques. Among these, the Louvain algorithm has become extremely popular due to its simplicity, efficiency and precision.

In this thesis will be presented an overview of community detection techniques and two new parallel implementations of the Louvain algorithm written in CUDA and exploitable by Nvidia GPUs: the first one is based on the sort-reduce paradigm with a pruning approach on the input data; the second one is a new hash-based implementation. Furthermore, it will be presented an analysis between these two approaches.

Contents

1 Introduction

2 Community Detection

The problem of community detection raised from the necessity of finding groups of objects that have a high number of connections to each other. To represent problems where it is fundamental to empathize connection between objects, the graph theory is the main tool. A graph is a mathematical structure composed by nodes (or vertices) that denote the objects, edges (or links) that express some kind of relationship between objects and eventually some weights that quantify this relationship.

The Graph Theory born in 1736 from Euler that used this mathematical abstraction to solve the puzzle of Königsberg's bridges. Since them, this tool was used in a lot of Mathematics, Social, Biological and Technological application. With the advent of Computer Science, the study of this field was has been revolutionized to deal with bigger and more complicated problems, supported by the new computing power. The necessity of finding this high-connected substructure in graph arises from real problems of the previous field: for example, the study of Protein-Protein Interaction (PPI) networks is very important because the interaction between proteins is the basis of all process in the cell. A study demonstrated that this type of network shown to be useful for highlighting key proteins involved in metastasis. [5]

Other examples can be found in the field of sociology: a well-know scenario is the Zachary's Karate Club. This dataset captures members of a Karate Club for 3 years.[3] An edge between two nodes represents an interaction between two members outside the club. At some point, a conflict between the administrator and a master led to split of the club into two separate groups. The question is if it is possible to infer who compose these two new group basing on the information that this graph give to us. In general this kind of problem, i.e. clustering people that belong to the same community base on interaction, it's useful not only in sociology but also in marketing. Knowing people with similar interest, it's possible to make better recommendation systems.

There is a lot of similar scenario in the real-word to apply this method. The key fact that unites all this situation is that aren't regular: "they are objects where order coexists with the disorder". [6] A fully disordered graph is the random graph, firstly proposed by Erdös and Rényi [1] in 1959, it's a graph where the probability that there is an edge between two nodes it's equal for all pairs of nodes and, for this reason, the degree of the nodes (i.e. the number of edges incident to a node) is homogeneous. In real networks, this is not true, because they are often scale-free (fallow a power-law distribution). An example of this is the study about the citations in scientific papers made by Derek J. de Solla Price in 1965 [2] or the study about World Wide Web growing made by Albert-Lászó Barabási et al in 1999 [4]. Furthermore, the distribution is inhomogeneity not only globally but also locally, this due to the observation that there is a high concentration of edges within sets of nodes and a low concentration of edges between this sets. These two concepts are essential to formulate the formal definition of Community and Modularity, that it will be presented in the following chapters.

2.1 Community Definitions

There isn't a unique quantitative definition of community. This kind of freedom is necessary because the concept of community is strictly connected to the problem that will be analyzed: for example, in some cases, it's necessary that community overlap, but in other problems, this is not necessary. There is a unique key constraint that allows talking about community detection: the graph must be sparse. A sparse graph is a graph where the number of nodes has the same magnitude of the number of edges. In the unweighted graph case, if the number of edges is far greater than the number of nodes, the distribution of edges among the nodes is too homogeneous for communities to make sense. [6]. In that case, the problem nature is little different: we aren't interested anymore on the edge density between nodes but we have to use some kind of metrics (like similarity or distance) to clustering. In that case, the problem is more similar to data clustering. Despite this, assuming that a community is a set of similar nodes it's reasonable, for this reasons some techniques (like spectral o hierarchical clustering) belonging to this field are adopted in community detection and will be shortly presented later on this thesis.

The main idea behind communities is there are many more edges inside the community versus the rest of the graph. Following this, Fortunato [6] defines three main classes of community's definitions: *local, global and based on vertex similarity.* Other types of definitions are still possible based on the situation, but these three offers give a good overview of the problem.

2.1.1 Local definitions

Given a graph $G(V, E)$ and a subset of nodes C where $C \in V$, we define the internal degree k_v^{int} and external degree k_v^{ext} for each vertex v that belongs to C as the number of edges that connect the node v with another node that belongs to C and not, respectively.

2.2 Community Detection Algorithms

3 Modularity Optimization

3.1 Sequential Algorithm

3.2 Parallel Algorithm

4 Nvidia GPUs architecture and CUDA

5 Implementation

5.1 Overview

5.2 Fast-Move Pruning

5.3 Sort-Reduce Pattern

6 Performance and Analysis

7 Conclusion

References

[1] A.Rényi P.Erdös. “On Random Graphs”. In: *Publ. Math. Debrecen 6* (Dec. 1959), pp. 290–297.

[2] Derek J. de Solla Price. “Networks of Scientific Papers”. In: *Science* 149.3683 (1965), pp. 510–515. ISSN: 0036-8075. DOI: 10.1126/science.149.3683.510. eprint: https://science.sciencemag.org/content/149/3683/510.full.pdf. URL: https://science.sciencemag.org/content/149/3683/510.

[3] W.W. Zachary. “An information flow model for conflict and fission in small groups”. In: *Journal of Anthropological Research* 33 (1977), pp. 452–473.

[4] Albert-László Barabási and Réka Albert. “Emergence of Scaling in Random Networks”. In: *Science* 286.5439 (1999), pp. 509–512. ISSN: 0036-8075. DOI: 10.1126/science.286.5439.509. eprint: https://science.sciencemag.org/content/286/5439/509.full.pdf. URL: https://science.sciencemag.org/content/286/5439/509.

[5] Pall F. Jonsson et al. “Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis”. In: *BMC Bioinformatics* 7.1 (Jan. 2006), p. 2. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-2. URL: https://doi.org/10.1186/1471-2105-7-2.

[6] Santo Fortunato. “Community detection in graphs”. In: *Physics Reports* 486.3-5 (Feb. 2010), pp. 75–174. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2009.11.002. URL: http://dx.doi.org/10.1016/j.physrep.2009.11.002.