Community Detection: Metaheuristic

Jeroen Craps
KU Leuven
jeroen.craps@student.kuleuven.be

Jorik De Waen
KU Leuven
jorik.dewaen@student.kuleuven.be

December 16, 2016

I. Introduction

Networks are ever so present in the world, due to the rise of social media and the emergence of Big Data[1] over the last decade. The detection of communities[2] in these large networks has grown in importance. Communities can be seen as fairly independent parts of the graph. The information contained in these graphs can be of utmost importance to understanding the data that is being dealt with. The problem has been proven to be NP-hard [4].

II. Overview

A large amount of research has gone into community detection in the last several decades. Finding a good way to approach this issue has been particularly hard because it is not trivial to come up with an exact definition of a community and a metric to compare different partitionings. As the amount of data being gathered grows at an incredible pace, so has the size of the networks which need to be analysed.

Traditional approaches simply don't scale to many thousands, let alone millions, of nodes. Because the problem is NP-hard, calculating the optimal solution for large networks is an unreachable goal. Genetic algorithms provide a way to still find solutions in such large networks. One such algorithm was introduced in a recent paper by Li et al. [5], using a multi-agent approach. Each agent represents a candidate solution and "lives" in a lattice structure. In this lattice, candidate solutions compete with their direct neighbours.

Most algorithms for community detection assume that each node can only be a part of a single community. In many cases, this is simply not true. Overlapping community detection algorithms can be divided into two groups: node-based algorithms and link-based algorithms [2].

The node-based algorithms focus directly on the nodes and try to detect communities by looking at how nodes are related. The link-based algorithms are built with the assumption that the links between nodes are actually more important than the nodes themselves. Not the individuals, but the relations between the individuals define the community. The links are divided into communities, and only afterwards is that translated to the nodes. Generally, link-based algorithms have been shown to yield superior results, but at a much higher computational cost. Ding et al. [6] have proposed a new approach which attempts to improve on the computational cost typically associated with a link-based algorithm using network decomposition. This algorithm is not genetic, but others have proposed several different genetic overlapping community detection algorithms [2, 3, 1].

[1] Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them.

[2] A network is said to have community structure if the nodes of the network can be easily grouped into (potentially overlapping) sets of nodes such that each set of nodes is densely connected internally.

III. Methods

Based on previous work on overlapping community detection algorithms, the intention of this research is to use the technique of node clustering[6] to improve the efficiency and performance of a recently used Multi-Agent community detection algorithm[5]. The candidate solutions in our genetic overlapping community detection algorithm will have the same structure as presented in the multi-agent algorithm by Li et al. We believe that some of the techniques used to detect overlapping algorithms could be applied to the multi-agent algorithm to allow it to detect overlapping communities as well, seeing that this was not the case up until now.

In the locus-based adjacency representation of the graph structure, all of the nodes are represented separately. This leads to a very large data structure when being used on large, but realistic network graphs. The space and time complexity decreases when the amount of nodes is decreased. A decrease in the amount of nodes is made by clustering together nodes that are clearly interconnected. This will be done by randomly selecting seeds in the network to looks for strong local communities. Due to the use of edge contraction, the general structure of the graph will remain without much information loss. That is if the relationship between the merged nodes in the graph is transitive. Multiple methods of link clustering will be tested to see their performances compared to the original set-up.

A problem that might arise is the fact that it might become harder to find nodes that are in several communities at the same time. Further research is required to see how this can be prevented, while still improving the efficiency of the Genetic Algorithm.

References

[1] Wei Hu Brian Dickinson, Benjamin Valyou. A genetic algorithm for identifying overlapping communities in social networks using an optimized search space. *Social Networking*, pages 193–201, 2013.

[2] Di Fu Yuxiao Dong Chuan Shi, Yanan Cai and Bin Wu. A link clustering based overlapping community detection algorithm. *Data & Knowledge Engineering*, 87:394–404, 2013.

[3] Pizzuti Clara. Overlapped community detection in complex networks. *GECCO*, pages 859–866, 2009.

[4] S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, Febuary 2010.

[5] Zhangtao Li and Jing Liu. A multi-agent genetic algorithm for community detection in complex networks. *Physica A*, 449:336–347, 2016.

[6] Dengdi Sun Zhuanlian Ding, Xingyi Zhang and Bin Luo. Overlapping community detection based on network decomposition. *Scientific Reports*, 2016.