Community Detection: Metaheuristic

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

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

December 10, 2016

I. Introduction

Communities are found in networks everywhere in the world, due to the rise of social networks and the collection of other graph data over the last decade these networks have been growing exponentially. Communities can be seen as fairly independent parts of the graph. The information these graphs contain can be very valuable for a variety of reasons. Faster algorithms are required to deal with this increase of data. 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 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].

DIT IS MISSCHIEN MEER METHODE? The candidate solutions in these genetic overlapping community detection algorithms have the same structure as the multi-agent algorithm by Li et al. We believe that some of the techniques used in the overlapping algorithms could be applied to the multi-agent algorithm to allow it to detect overlapping communities.

III. Methods

Based on previous algorithms that have been used detect overlapping communities in large networks, the intention is to use techniques from outside of the Genetic Algorithm domain to improve the efficiency and performance of a recently used Multi-Agent approach [5]. The main method that will be used to facilitate is Node Clustering [6]. In the locus-based adjacency representation of the graph structure, all of the nodes are represented separately which leads to a very large data structure when being used on large but realistic network graphs. By clustering some of the obvious communities into new nodes prior to using the Genetic Algorithm, the space and time complexity should decrease. Edge contraction based on node clustering should keep the general structure of the graph without much information loss if the relationship between the merged nodes in the graph is transitive. Multiple methods of link clustering will be tested to see its performance compared to the original set-ups A possible problem might be that by doing this is will become harder to find nodes that are in several communities. 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.