

# **Overlapping community detection in complex networks with memetic algorithms**

by

**Ademir Cristiano Gabardo**

**Ph.D. Thesis**

*submitted in partial fulfilment  
of the requirements for the Degree of*

**Doctor of Philosophy**

Supervisor: Professor Regina Berretta

Co-supervisor: Professor Pablo Moscato



THE UNIVERSITY OF  
**NEWCASTLE**  
AUSTRALIA

**The University of Newcastle**

**Faculty of Engineering and Built Environment  
School of Electrical Engineering and Computing**

July, 2018



© Copyright

by

Ademir Cristiano Gabardo

2018



# **Overlapping community detection in complex networks with memetic algorithms**

## **Statement of Originality**

I hereby certify that the work embodied in the thesis is my own work, conducted under normal supervision.

The thesis contains no material which has been accepted, or is being examined, for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to the final version of my thesis being made available worldwide when deposited in the University's Digital Repository, subject to the provisions of the Copyright Act 1968 and any approved embargo.

Ademir Cristiano Gabardo

31st July 2018



# Acknowledgments

First, I would like to express my sincere gratitude to my advisers, Professor Regina Berretta and Professor Pablo Moscato, for their valuable knowledge and continuous support of my PhD study. Their guidance helped me throughout my research and while writing this thesis. I could not have imagined having better advisers for my PhD.

This research was supported by the Brazilian National Council for Scientific and Technological Development (CNPQ).<sup>1</sup>

A research thesis may have a single author. However, this thesis is unquestionably not a journey of a single person there is so much to consider and there are so many to thank. These brief pages are not enough to express the enormous gratitude I have for all those somehow affected by my decision to pursue a PhD. With love and passion, I acknowledge my wife, ♡Juliana Gabardo♡, who supported me during this long journey. Without her love, I certainly would not have survived the challenging paths of academic research.

I also acknowledge my mother, Maria Francisca Gabardo, for her enormous courage in crossing the oceans and coming to Australia to bring me her heartwarming mother's love, and my father, Jair Gabardo, for supporting my decision and encouraging me to move forward. With deep respect, I acknowledge my father-in-law and my mother-in-law, who promptly supported our decision to move abroad.

I acknowledge my son, Adrian Gabardo, for embarking on this great adventure by my side and endure the challenges of moving overseas.

I am certain that the most valuable treasure in life are the good friends we meet along the way, and acknowledge my friends from the Priority Research Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine and Life and Economy Applications of Data Science research groups: Amer Abu Zaher, Amir

---

<sup>1</sup>CNPq - <http://cnpq.br>

Salehipour, Claudio Sanhueza Lobos, Francia Jimenez, Heloisa Milioli, Jake Fitzsimmons, Inna Tischenko, Leila Moslemi Naeni, Luke Mathieson, Lukasz Olech, Marta A. Jimenez, Nader Mahmoudi, Natalie de Vries, Nasimul Noman and Shannon Fenn. I thank them for the friendly environment, the countless coffee breaks and the support.

Although distant, but always present, I acknowledge my friend and mentor, Professor Heitor Silvério Lopes, who always had a word of support and encouragement.

I also extend my deep respect and esteem to the other friends have also been part of this journey, especially Luiz Fernando Nunes, Lauro Cesar Galvão, Chidambaram Chidambaram and Osvalcir Peters.

*Ademir Cristiano Gabardo*

*The University of Newcastle*

*July 2018*

# Contents

|                                                              |             |
|--------------------------------------------------------------|-------------|
| <b>Acknowledgments</b> . . . . .                             | <b>vii</b>  |
| <b>List of Tables</b> . . . . .                              | <b>xiii</b> |
| <b>List of Figures</b> . . . . .                             | <b>xvii</b> |
| <b>Abstract</b> . . . . .                                    | <b>xxvi</b> |
| <b>1 Introduction</b> . . . . .                              | <b>1</b>    |
| 1.1 Networks and applications . . . . .                      | 1           |
| 1.1.1 Community structures in networks . . . . .             | 4           |
| 1.1.2 Overlapping community structures in networks . . . . . | 5           |
| 1.2 Research problem . . . . .                               | 6           |
| 1.3 Research objectives . . . . .                            | 7           |
| 1.4 Organisation of the thesis . . . . .                     | 8           |
| 1.5 List of publications and presentations . . . . .         | 9           |
| <b>2 Concepts and definitions</b> . . . . .                  | <b>11</b>   |
| 2.1 Networks and graphs . . . . .                            | 11          |
| 2.2 Graph properties . . . . .                               | 14          |
| 2.2.1 Node degree . . . . .                                  | 14          |
| 2.2.2 Degree distribution . . . . .                          | 15          |
| 2.2.3 Centrality measures . . . . .                          | 15          |
| 2.2.4 Node neighbourhood . . . . .                           | 19          |
| 2.2.5 Clique . . . . .                                       | 19          |
| 2.2.6 Connected components . . . . .                         | 20          |
| 2.2.7 Clustering coefficient . . . . .                       | 20          |
| 2.2.8 Graph density . . . . .                                | 21          |
| 2.3 Line graph . . . . .                                     | 22          |
| 2.4 Network models . . . . .                                 | 23          |
| 2.4.1 Erdős and Rényi model and random graphs . . . . .      | 23          |

|          |                                                                                    |           |
|----------|------------------------------------------------------------------------------------|-----------|
| 2.4.2    | Watts and Strogatz small-world model and the ‘six degrees of separation’ . . . . . | 24        |
| 2.4.3    | Barabási and Albert model and scale-free networks . . . . .                        | 26        |
| 2.5      | The community detection problem . . . . .                                          | 28        |
| 2.5.1    | Disjoint communities . . . . .                                                     | 29        |
| 2.5.2    | Overlapping communities . . . . .                                                  | 31        |
| 2.6      | Metrics for evaluating community quality . . . . .                                 | 32        |
| 2.6.1    | Modularity . . . . .                                                               | 32        |
| 2.6.2    | Link partition density . . . . .                                                   | 34        |
| 2.6.3    | Other metrics for community detection . . . . .                                    | 35        |
| 2.7      | Summary . . . . .                                                                  | 37        |
| <b>3</b> | <b>A review of methods and benchmarks for community detection . . . . .</b>        | <b>39</b> |
| 3.1      | Methods for overlapping community detection . . . . .                              | 39        |
| 3.1.1    | Clique-based methods . . . . .                                                     | 40        |
| 3.1.2    | Local-expansion methods . . . . .                                                  | 42        |
| 3.1.3    | Label-propagation methods . . . . .                                                | 46        |
| 3.1.4    | Link-clustering methods . . . . .                                                  | 48        |
| 3.1.5    | Other strategies for overlapping community detection . . . . .                     | 52        |
| 3.2      | Metaheuristics for community detection . . . . .                                   | 52        |
| 3.2.1    | Genetic algorithms for community detection . . . . .                               | 53        |
| 3.2.2    | Memetic algorithms for community detection . . . . .                               | 59        |
| 3.2.3    | Other metaheuristic approaches . . . . .                                           | 63        |
| 3.3      | Benchmarks for community detection . . . . .                                       | 66        |
| 3.3.1    | Real-world benchmark networks for community detection . . . . .                    | 66        |
| 3.3.2    | The Girvan and Newman benchmark . . . . .                                          | 68        |
| 3.3.3    | LFR synthetically generated benchmark networks . . . . .                           | 68        |
| 3.4      | Summary . . . . .                                                                  | 71        |
| <b>4</b> | <b>MADOC: A memetic algorithm to detect overlapping communities . . . . .</b>      | <b>73</b> |
| 4.1      | The general framework of MADOC . . . . .                                           | 74        |
| 4.2      | Line graphs for link-clustering . . . . .                                          | 75        |
| 4.3      | Individual representation . . . . .                                                | 75        |
| 4.4      | A multi-agent ternary tree population . . . . .                                    | 76        |
| 4.5      | Population Initialisation . . . . .                                                | 78        |
| 4.5.1    | Creating individuals with local-expansion . . . . .                                | 78        |
| 4.5.2    | Creating individuals with label-propagation . . . . .                              | 80        |
| 4.6      | Genetic operators . . . . .                                                        | 82        |
| 4.6.1    | Modularity recombination crossover . . . . .                                       | 82        |
| 4.6.2    | Neighbouring communities mutation . . . . .                                        | 83        |

|          |                                                                                                         |            |
|----------|---------------------------------------------------------------------------------------------------------|------------|
| 4.7      | Selection and replacement . . . . .                                                                     | 85         |
| 4.8      | Local-search algorithms . . . . .                                                                       | 86         |
| 4.8.1    | A local-search algorithm for modularity maximisation . . . . .                                          | 86         |
| 4.8.2    | A deterministic-update annealing for modularity<br>maximisation . . . . .                               | 87         |
| 4.9      | Computational experiments and results . . . . .                                                         | 89         |
| 4.9.1    | Computational environment and parameters . . . . .                                                      | 89         |
| 4.9.2    | A comparison between the hill-climbing local-search and the<br>deterministic-update annealing . . . . . | 90         |
| 4.9.3    | Modularity maximisation results . . . . .                                                               | 91         |
| 4.9.4    | Experiments in synthetically generated overlapping<br>benchmark networks . . . . .                      | 93         |
| 4.10     | Conclusion and remarks . . . . .                                                                        | 95         |
| <b>5</b> | <b>Co-purchases for luxury brands at Amazon.com: A case study . . . . .</b>                             | <b>97</b>  |
| 5.1      | Introduction . . . . .                                                                                  | 97         |
| 5.2      | Constructing the network . . . . .                                                                      | 98         |
| 5.3      | Modularity maximization over the line graph . . . . .                                                   | 100        |
| 5.4      | Results and discussion . . . . .                                                                        | 101        |
| 5.5      | Conclusion . . . . .                                                                                    | 106        |
| <b>6</b> | <b>Co-purchases of photographic products: A case study . . . . .</b>                                    | <b>107</b> |
| 6.1      | Introduction . . . . .                                                                                  | 107        |
| 6.2      | Co-purchasing networks . . . . .                                                                        | 108        |
| 6.3      | Constructing the network . . . . .                                                                      | 109        |
| 6.4      | Modularity maximization over the line graph . . . . .                                                   | 111        |
| 6.5      | Results and discussion . . . . .                                                                        | 111        |
| 6.6      | Conclusion . . . . .                                                                                    | 116        |
| <b>7</b> | <b>MEMELink: A memetic algorithm for link partition density maximisation</b>                            | <b>117</b> |
| 7.1      | The general framework of MEMELink . . . . .                                                             | 117        |
| 7.2      | Individual representation . . . . .                                                                     | 118        |
| 7.3      | A multi-agent ternary tree population . . . . .                                                         | 119        |
| 7.4      | Population initialisation . . . . .                                                                     | 120        |
| 7.4.1    | Creating individuals with local-expansion . . . . .                                                     | 121        |
| 7.4.2    | Creating individuals with label-propagation . . . . .                                                   | 121        |
| 7.4.3    | Eigenvector centrality . . . . .                                                                        | 122        |
| 7.5      | Genetic operators . . . . .                                                                             | 123        |
| 7.5.1    | Uniform crossover . . . . .                                                                             | 123        |
| 7.5.2    | Mutation . . . . .                                                                                      | 124        |

|          |                                                                                                    |            |
|----------|----------------------------------------------------------------------------------------------------|------------|
| 7.6      | Local-search . . . . .                                                                             | 124        |
| 7.7      | Post-processing . . . . .                                                                          | 125        |
| 7.8      | Computational experiments and results . . . . .                                                    | 127        |
| 7.8.1    | Computational environment and parameters . . . . .                                                 | 127        |
| 7.8.2    | Instances of the LFR synthetically generated benchmark networks . . . . .                          | 128        |
| 7.8.3    | A comparison between MADOC and MEMELink . . . . .                                                  | 129        |
| 7.8.4    | The effect of network size ( $n$ ) on NMI and time consumption . . . . .                           | 131        |
| 7.8.5    | The effect of mixing parameter $\mu$ and number of community assignments ( $om$ ) on NMI . . . . . | 133        |
| 7.8.6    | The effect of overlapping extension ( $on$ ) and community assignments ( $om$ ) on NMI . . . . .   | 136        |
| 7.8.7    | Comparing MEMELink with other algorithms . . . . .                                                 | 137        |
| 7.9      | Conclusion and remarks . . . . .                                                                   | 139        |
| <b>8</b> | <b>A storm of swords: A case study . . . . .</b>                                                   | <b>141</b> |
| 8.1      | Introduction . . . . .                                                                             | 141        |
| 8.2      | The network of characters in the novel ‘A storm of swords’ . . . . .                               | 142        |
| 8.3      | Results of MADOC and discussion . . . . .                                                          | 143        |
| 8.4      | Results of MEMELink and discussion . . . . .                                                       | 147        |
| 8.5      | Conclusion . . . . .                                                                               | 150        |
| <b>9</b> | <b>Conclusions and future work . . . . .</b>                                                       | <b>151</b> |
| 9.1      | Conclusions . . . . .                                                                              | 151        |
| 9.1.1    | MADOC . . . . .                                                                                    | 151        |
| 9.1.2    | MEMELink . . . . .                                                                                 | 152        |
| 9.1.3    | Overlapping communities of luxury brands . . . . .                                                 | 153        |
| 9.1.4    | Overlapping communities from a brand-centric point of view . . . . .                               | 153        |
| 9.1.5    | Overlapping communities of characters in a novel . . . . .                                         | 154        |
| 9.2      | Future work . . . . .                                                                              | 154        |
|          | <b>Bibliography . . . . .</b>                                                                      | <b>157</b> |

# List of Tables

|     |                                                                                                                                                                                                                                                                                                                                                                      |    |
|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.1 | A symmetric adjacency matrix representing the graph in Figure 2.3. . . . .                                                                                                                                                                                                                                                                                           | 13 |
| 3.1 | Algorithms that employ cliques for overlapping community detection. Where $h_{max}$ is the number of number of pairs of maximal cliques that are neighbours, $s_{max}$ is the number of maximal cliques, and $ C $ is the number of cliques. . . . .                                                                                                                 | 41 |
| 3.2 | Seed-expansion and local-expansion algorithms for overlapping community detection where $m$ represents the number of edges, $n$ represents the number of nodes and $s$ is the number of communities. . . . .                                                                                                                                                         | 45 |
| 3.3 | Label-propagation algorithms for overlapping community detection where $t$ is the number of iterations. . . . .                                                                                                                                                                                                                                                      | 47 |
| 3.4 | Link-clustering algorithms for overlapping community detection where $t$ is the number of iterations, $s$ is the number of communities, $k_{max}$ is the maximum degree of a node in the graph, $n$ is the number of nodes, $m$ is the number of edges in a graph and $p_s$ is the population size. . . . .                                                          | 50 |
| 3.5 | Genetic algorithms for disjoint and overlapping community detection ordered by date of publication. . . . .                                                                                                                                                                                                                                                          | 56 |
| 3.6 | Memetic algorithms for disjoint and overlapping community detection ordered by year of publication or submission. . . . .                                                                                                                                                                                                                                            | 60 |
| 3.7 | Bio-inspired and other approaches metaheuristics for disjoint and overlapping community detection. . . . .                                                                                                                                                                                                                                                           | 64 |
| 3.8 | Real-world networks frequently used to assess the efficiency of community detection algorithms. . . . .                                                                                                                                                                                                                                                              | 67 |
| 4.1 | Comparison results for the minimum, average and maximum modularity between the local-search (LS) algorithm and the deterministic-update annealing (SA) for 50 individual runs for eight distinct benchmark networks. Columns show the benchmark, the minimum, average, maximum and the best-known modularity for both algorithms and the standard deviation. . . . . | 90 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                             |     |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.2 | Minimum, average, maximum and the standard deviation of the time in seconds spent by local-search and deterministic-update annealing for a single run in one individual. . . . .                                                                                                                                                                                                                                                                            | 91  |
| 4.3 | Experimental results for seven real-world benchmark networks showing the maximum and average modularity $Q$ and the standard deviation obtained by five algorithms for modularity maximisation and the best-known modularity for each benchmark. . . . .                                                                                                                                                                                                    | 92  |
| 4.4 | The average time spent in seconds and the standard deviation for 50 executions of MADOC in real-world benchmark networks. . . . .                                                                                                                                                                                                                                                                                                                           | 93  |
| 4.5 | Synthetically generated networks used to assess MADOC's efficiency. Columns show the network, number of nodes and edges, mixing parameter $\mu$ , number of overlapping nodes ( $on$ ), number of community assignments for the overlapping nodes ( $om$ ) and the number of communities $\#C$ . The table also shows the number of nodes and edges in the corresponding line graph $L(G)$ and the maximum modularity produced by MADOC in $L(G)$ . . . . . | 94  |
| 5.1 | Network statistics for the graph $G$ and its respective line graph $L(G)$ .                                                                                                                                                                                                                                                                                                                                                                                 | 100 |
| 5.2 | Product count and percent coverage for the overlapping communities C1 to C7 recovered from $L(G)$ to $G$ . . . . .                                                                                                                                                                                                                                                                                                                                          | 100 |
| 5.3 | Distribution of categories and gender according to the number of communities in which the products are assigned. The numbers at the left of the columns represent the product count, the targeted gender for the products are identified as $M$ for male, $F$ for female and $U$ for unisex. The categories are ordered according to the number of products. . . . .                                                                                        | 102 |
| 5.4 | The top 10 brands with the highest frequency among communities C1 to C7. (For short, in this table, Victoria's Secret = V.S., Dolce & Gabbana = D&G, Paris Hilton = P.H., Carolina Herrera = C.H., Calvin Klein = C.K, Paco Rabanne = P.R. and Emporio Armani = E.A.) . . . . .                                                                                                                                                                             | 103 |
| 5.5 | Columns indicate: Community A, its size, the percentage of nodes which overlaps with Community B, Community B size, the percentage of nodes which overlaps with Community A and, the number of overlapping nodes between Communities A and B. . . . .                                                                                                                                                                                                       | 104 |
| 6.1 | Network statistics for the graph $G$ and its respective line graph $L(G)$ .                                                                                                                                                                                                                                                                                                                                                                                 | 111 |
| 6.2 | The number of products, the minimum, the average, the maximum price and the standard deviation for products in communities C1 to C11 for the Kodak co-purchasing network. . . . .                                                                                                                                                                                                                                                                           | 111 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |     |
|-----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 6.3 | The brands with largest number of products in Kodak's co-purchasing network. . . . .                                                                                                                                                                                                                                                                                                                                                                                                              | 113 |
| 7.1 | The gain in NMI achieved with the post-processing for the benchmark networks LFR1 to LFR6. . . . .                                                                                                                                                                                                                                                                                                                                                                                                | 126 |
| 7.2 | LFR benchmark networks with 128 to 5,000 nodes, mixing parameter $\mu = 0.1$ or $0.3$ , number of overlapping nodes corresponding to approximately 10% or 20% of the number of nodes in the network and community assignments ( $om$ ) varying from 2 to 8. $\#C$ indicates the number of communities. . . . .                                                                                                                                                                                    | 129 |
| 7.3 | Comparative results between MADOC and MEMELink in a collection of 21 benchmark networks detailed in Table 7.2. Columns show the network, the average and maximum NMI and the standard deviation for both algorithms. . . . .                                                                                                                                                                                                                                                                      | 130 |
| 7.4 | Comparative results for the algorithms MADOC, SLPA and MEMELink for the benchmark networks LFR1 to LFR6, shown in Table 7.2. Columns show the network and the average, maximum and the standard deviation of NMI for both algorithms. We also report the average and maximum link partition density $D$ and the standard deviation achieved by MEMELink. . . . .                                                                                                                                  | 132 |
| 7.5 | Average time in seconds required to detect the overlapping communities in the benchmark networks LFR1 to LFR6 for the algorithms SLPA, MEMELink and MADOC. . . . .                                                                                                                                                                                                                                                                                                                                | 133 |
| 7.6 | Comparative results of MEMELink and SLPA for the networks LFR7 to LFR34 detailed in Table 7.2. Columns show the network and the average, maximum and the standard deviation of link partition density $D$ achieved by MEMELink and, the average, maximum and the standard deviation of NMI achieved by MEMELink and SLPA. . . . .                                                                                                                                                                 | 135 |
| 7.7 | Comparative results of MEMELink and SLPA for the networks LFR14 to LFR20 and LFR35 to LFR41 shown in Table 7.2, varying the number of overlapping nodes ( $on$ ) and the number of community assignments ( $om$ ) for the overlapping nodes. Columns show the network and the average, the maximum and the standard deviation for the link partition density $D$ achieved by MEMELink, and the average, the maximum and the standard deviation for the NMI achieved by SLPA and MEMELink. . . . . | 136 |
| 8.1 | The Pearson's, Spearman's and Kendall's correlation between the number of communities to which a character is associated with and metrics of strength and centrality. . . . .                                                                                                                                                                                                                                                                                                                     | 145 |

|     |                                                                                                                                         |     |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------|-----|
| 8.2 | The most central characters of the network “Storm of Swords” ordered according to the number of communities in which they appear. . . . | 146 |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------|-----|

# List of Figures

|     |                                                                                                                                                                                                                                                                                                                                                                                                                   |    |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1 | (Left), a 3D model of the worm <i>Caenorhabditis Elegans</i> . (Right) The network of protein-encoding genes of the worm <i>Caenorhabditis elegans</i> [257]. The colours represent the communities detected by our memetic algorithm presented in Chapter 4, showing groups of genes that interact with one another. . . . .                                                                                     | 2  |
| 1.2 | The network of 500 airports with the largest amount of traffic in the United States, the connections represent the traffic of passengers between airports. The size of the nodes represents the amount of traffic in the airport, and the colours represent the communities detected by our memetic algorithm presented in Chapter 4, showing groups of airports with intense traffic between each other. . . . . | 3  |
| 1.3 | A network of 128 nodes divided into four communities represented by distinct colours and node shapes showing a higher density of connections inside communities and few connections between communities. The network is a modified instance of the Girvan and Newman benchmark presented in Section 3.3.2, and is edited and coloured with the software yEd graph editor. . . . .                                 | 4  |
| 1.4 | Network with 13 nodes organised in two overlapping communities. The network shows groups of students affiliated to different scholar clubs. Nodes six and seven are assigned to both communities, and are therefore overlapping. Network and figure are generated with the software yEd graph editor. . . . .                                                                                                     | 6  |
| 2.1 | The seven bridges of Königsberg are represented by <i>a</i> to <i>g</i> ; the island of <i>Kneiphof</i> is represented by <i>A</i> , and <i>B</i> , <i>C</i> and <i>D</i> , showing distinct portions of land divided by the Pregel River. The lines connecting <i>A</i> , <i>B</i> , <i>C</i> and <i>D</i> are the graph representation of the problem Euler proposed. . . . .                                   | 12 |
| 2.2 | Adapted from: The elements of network models - Brandes, Ulrik, et al. 'What is network science?' Network Science 1.01 (2013): 1-15. . . .                                                                                                                                                                                                                                                                         | 12 |

|      |                                                                                                                                                                                                                                                                                                                                                                                       |    |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.3  | (Left) A graph of six nodes joined by seven edges. (Right) The set of nodes and edges for the graph on the left. . . . .                                                                                                                                                                                                                                                              | 12 |
| 2.4  | a) An unweighted undirected graph. b) A weighted undirected graph. c) An unweighted directed graph. d) A weighted directed graph. e) A mixed graph with undirected and directed edges, parallel edges and self-loop. . . . .                                                                                                                                                          | 13 |
| 2.5  | The adjacency list corresponding to the graph of six nodes and seven edges from Figure 2.3. . . . .                                                                                                                                                                                                                                                                                   | 14 |
| 2.6  | (Left) The degree distribution of a random graph; the trend line shows the binomial distribution. (Right) The degree distribution of a real-world network representing the 500 airports with the largest amount of traffic in the US; the trend line shows a power-law distribution, where a few nodes have a high number of connections and most nodes have few connections. . . . . | 16 |
| 2.7  | A graph with nine nodes and 15 edges where node 5 is the most central node connecting the two groups of nodes highlighted by distinct colours. . . . .                                                                                                                                                                                                                                | 16 |
| 2.8  | A graph of six nodes joined by seven edges with the neighbourhood of node 3 comprising nodes $\{2, 3, 4, 5\}$ , highlighted in yellow. . . . .                                                                                                                                                                                                                                        | 19 |
| 2.9  | A complete graph with five nodes in which every node connects to the other nodes in the network, forming a clique. . . . .                                                                                                                                                                                                                                                            | 19 |
| 2.10 | A graph with 11 nodes and 12 edges showing three connected components. . . . .                                                                                                                                                                                                                                                                                                        | 20 |
| 2.11 | (Left) A sparse graph with 12 nodes and 21 edges with graph density $D(G) = 0.318$ . (Centre) A dense graph with 12 nodes and 43 edges with graph density $D(G) = 0.652$ . (Right) A completely connected graph with 12 nodes and 66 edges with graph density $D(G) = 1$ . . . . .                                                                                                    | 21 |
| 2.12 | (Left) A graph $G$ with eight nodes and 16 edges. (Right) The corresponding line graph $L(G)$ of $G$ with 16 nodes and 54 edges. The edges between the nodes 1, 5 and 8 are highlighted in colour in the graph, and the corresponding nodes 1,5 and 5,8, and the edge between these nodes are highlighted in colour in the line graph. . . . .                                        | 22 |
| 2.13 | (Left) A graph with 79 nodes and 217 edges with average degree $\langle k \rangle = 5.494$ generated with the Erdős and Rényi model. (Right) The bar graph of its degree distribution showing a <i>normal distribution</i> where the majority of nodes have a degree similar to the average degree of the graph. . . . .                                                              | 24 |

|      |                                                                                                                                                                                                                                                                                                                                                           |    |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.14 | (Left) A graph with 75 nodes and 75 edges and average degree $\langle k \rangle = 2$ generated with the Watts and Strogatz model. (Right) The respective degree distribution for this graph. The majority of the nodes have the degree similar to the average degree in the network, and few nodes have a higher degree connecting to many nodes. . . . . | 25 |
| 2.15 | (Left) An example of a small-world graph with complex interconnections. (Right) The group of $X$ 's acquaintances feedback into his own circle, normally eliminating new contacts. Figure adapted from; Travers, Jeffrey, and Stanley Milgram. 'The small-world problem', <i>Psychology Today</i> 1 (1967): 61–67. . . . .                                | 26 |
| 2.16 | (Left) A graph with 77 nodes and 84 edges and average degree $\langle k \rangle = 1.091$ generated according to the Barabási and Albert model. (Right) The respective degree distribution following a power law distribution $P(k) \sim k^{-3}$ . . . . .                                                                                                 | 27 |
| 2.17 | a) A network divided into three disjoint communities. b) A network divided into three overlapping communities. c) A network divided into two hierarchical communities. d) A network divided into three soft overlapping communities. . . . .                                                                                                              | 29 |
| 2.18 | A network with 19 nodes divided into four disjoint communities. . . .                                                                                                                                                                                                                                                                                     | 30 |
| 2.19 | A network with 15 nodes divided into three overlapping communities. . .                                                                                                                                                                                                                                                                                   | 31 |
| 3.1  | A timeline highlighting the appearance of the major contributions in the overlapping community detection over the years: clique-based, local-expansion, label-propagation and link-clustering methods for overlapping community detection. . . . .                                                                                                        | 39 |
| 3.2  | Two overlapping communities comprising two five-node cliques sharing four nodes. . . . .                                                                                                                                                                                                                                                                  | 41 |
| 3.3  | The closed neighbourhood of the node '1' comprising by the node, its neighbours and the edges connecting this set of nodes. . . . .                                                                                                                                                                                                                       | 44 |
| 3.4  | A small network after the first interaction of the algorithm. Each node shows the chance that it has to receive the label of its neighbour. Figure adapted from 'Finding overlapping communities in networks by label propagation' by Steve Gregory, 2010, <i>New Journal of Physics</i> , 103018, v.12, n.10. . . . .                                    | 46 |
| 3.5  | (Left) A graph $G$ with eight nodes and 16 edges. (Centre) The line graph $L(G)$ of $G$ with 16 nodes and 54 edges partitioned into two communities. (Right) The graph $G$ with the communities found in $L(G)$ highlighted in colour; node 5 appears in both communities in $L(G)$ . Therefore, node five is assigned to both communities in $G$ . . .   | 49 |

|     |                                                                                                                                                                                                                                                                                                                                                                                   |    |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.6 | A network of 10 nodes and 25 edges divided into two edge communities. Nodes are assigned to communities of their adjacent edges. Node six is adjacent to edges in both communities. Therefore, node six is an overlapping node. . . . .                                                                                                                                           | 50 |
| 3.7 | (Left) An example of the Girvan and Newman benchmark network divided into four communities highlighted by colour. (Right) At the top, a line graph showing the degree distribution; all nodes have the same degree $k = 16$ . At the bottom is a bar chart showing the community size distribution—the four communities are of equal size with 32 nodes. . . . .                  | 69 |
| 3.8 | (Left) A LFR benchmark network with the following parameters: $n = 500$ , $\tau_1 = 1$ , $\tau_2 = 2$ , $\langle k \rangle = 10$ , $\max \langle k \rangle = 25$ and $\mu = 0.1$ . (Right) On top, the degree distribution for this network. Right, at the bottom, the community size distribution for this network. The communities are highlighted by distinct colours. . . . . | 70 |
| 4.1 | (Left) A graph $G = (V, E)$ . (Centre) The line graph $L(G)$ of $G$ divided into two communities that obtained the maximum modularity highlighted by colours. (Right) The overlapping communities of nodes in $G$ recovered from the edge communities. . . . .                                                                                                                    | 75 |
| 4.2 | (Left) An example of a graph $G$ with seven nodes and 12 edges. (Centre) The corresponding $L(G)$ of $G$ , with 12 nodes and 33 edges. In parenthesis, we kept the labels of the nodes corresponding to the edges in $G$ . (Right) The string-coding representation of the individual for $L(G)$ ; the colours of nodes represent the disjoint communities of $L(G)$ . . .        | 76 |
| 4.3 | The multi-agent tertiary tree population employed by MADOC. . . . .                                                                                                                                                                                                                                                                                                               | 77 |
| 4.4 | Communities of nodes generated by the node seed expansion method shown in Algorithm 4.2 for the Zachary karate club network. Square nodes represent seeds; the distinct colours highlight the community structure. . . . .                                                                                                                                                        | 79 |
| 4.5 | Communities of nodes generated by the label propagation method shown in Algorithm 4.3 for the Zachary karate club network. Square nodes represent seeds; the distinct colours highlight the community structure. . . . .                                                                                                                                                          | 81 |
| 4.6 | Two individuals ( <i>parent1</i> and <i>parent2</i> ) representing distinct community structures for a network recombined as a new individual ( <i>offspring</i> ) by the modularity-based recombination crossover. . . . .                                                                                                                                                       | 83 |
| 4.7 | A network divided into four communities labelled as C1, C2, C3 and C4 and highlighted by distinct colours. . . . .                                                                                                                                                                                                                                                                | 84 |

|     |                                                                                                                                                                                                                                                                                                              |     |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.8 | Examples of the <i>Close</i> and <i>Far</i> selection mechanisms employed by MADOC. . . . .                                                                                                                                                                                                                  | 85  |
| 5.1 | A record of the database showing the product 'Pasha De Cartier By Cartier For Men' stored within the MongoDB document structure and its respective fields. . . . .                                                                                                                                           | 98  |
| 5.2 | The distribution gender and categories on the Versace co-purchasing network. . . . .                                                                                                                                                                                                                         | 99  |
| 5.3 | Community size distribution for communities C1 to C7. . . . .                                                                                                                                                                                                                                                | 100 |
| 5.4 | Communities of products C1 to C7 in the co-purchasing network highlighted by distinct colors. . . . .                                                                                                                                                                                                        | 101 |
| 5.5 | Distribution of nodes according to the number of communities they participate. . . . .                                                                                                                                                                                                                       | 101 |
| 5.6 | The predominant category of products and the targeted gender of items for communities C1 to C7. . . . .                                                                                                                                                                                                      | 103 |
| 5.7 | The network of brands of co-purchases of luxury items on Amazon.com. The weight of the edges, shown by its thickness, represents the number of co-purchases between two brands. The network is divided into four communities ( <i>a</i> , <i>b</i> , <i>c</i> and <i>d</i> ) highlighted by colours. . . . . | 105 |
| 6.1 | A sample record of the database showing the product 'Kodak Ink Jet Photo Stickers' stored within the MongoDB document structure and its respective fields. . . . .                                                                                                                                           | 110 |
| 6.2 | (Left) Community size distribution for Kodak's co-purchasing network. (Right) The average, the minimum and the maximum price for products in USD \$ for communities C1 to C11 in the Kodak co-purchasing network. . . . .                                                                                    | 112 |
| 6.3 | Communities C1 to C11 highlighted by distinct colours in the Kodak co-purchasing network of photographic material. . . . .                                                                                                                                                                                   | 112 |
| 6.4 | (Left) The communities C2 and C4 highlighted by colours with the overlapping nodes highlighted by a darker colour. (Centre) The word-cloud of products found in the overlapping section. (Right) The word-cloud of brands in the overlapping intersection of communities C2 and C4. . . . .                  | 113 |
| 6.5 | (Left) The communities C11 and C4 highlighted by colours with the overlapping nodes highlighted by a darker colour. (Centre) The word-cloud of products found in the overlapping section. (Right) The word-cloud of brands in the overlapping intersection of communities C11 and C4. . . . .                | 114 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |     |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 6.6 | (Left) The communities C1 and C11 highlighted by colours with the overlapping nodes highlighted by a darker colour. (Centre) The word-cloud of products found in the overlapping section. (Right) The word-cloud of brands in the overlapping intersection of communities C1 and C11. . . . .                                                                                                                                                                                        | 115 |
| 6.7 | (Left) The community C7 highlighted by colour. (Centre) The word-cloud of products in this community. (Right) The word-cloud of brands in the community C7. . . . .                                                                                                                                                                                                                                                                                                                  | 115 |
| 6.8 | A detailed look at some nodes that are in a specific overlapping location. (Left) The central node clearly belongs to both communities, the algorithm was able to depict this structure while preserving the community structure. (Centre) The central node is overlapping different communities, where it shares several edges. (Right) the nodes sharing connections with other communities are overlapping, while the small community on the right is also depicted. . . . .      | 116 |
| 7.1 | (Left) A network comprised of 14 edges with two link communities <i>a</i> ) and <i>b</i> ) where node 4 is the overlapping node. (Right) The corresponding genotype using the LAR. . . . .                                                                                                                                                                                                                                                                                           | 119 |
| 7.2 | A tertiary tree of 13 agents, comprising a pocket solution and six current solutions each. . . . .                                                                                                                                                                                                                                                                                                                                                                                   | 120 |
| 7.3 | Output of the initialisation methods used by MEMELink; distinct colours indicate the community of the edges. (Left) An output of the local-expansion method translated to the communities of edges for the <i>Les Miserables</i> benchmark network. (Centre) An output of the label-propagation method for the same network. (Right) The output of the local-expansion with most central nodes used as seeds. The size of the nodes indicates their eigenvector centrality . . . . . | 122 |
| 7.4 | A small network with two well-defined, non-overlapping communities, <i>a</i> and <i>b</i> , connected by a bridging edge. . . . .                                                                                                                                                                                                                                                                                                                                                    | 125 |
| 7.5 | The effect of the number of overlapping nodes ( <i>on</i> ), number of community assignments ( <i>om</i> ) and the mixing parameter $\mu$ over the NMI. . . . .                                                                                                                                                                                                                                                                                                                      | 131 |
| 7.6 | (Left) The maximum and average NMI achieved for networks LFR1 to LRF6 with 128 up to 4,096 nodes. (Right) The average time in seconds for the same networks. . . . .                                                                                                                                                                                                                                                                                                                 | 133 |
| 7.7 | A comparison between the results of MEMELink and the SLPA algorithm for the four groups of benchmark networks, varying the mixing parameter $\mu$ and the number of community assignments for the overlapping nodes ( <i>om</i> ). . . . .                                                                                                                                                                                                                                           | 135 |

|     |                                                                                                                                                                                                                                                                                                                                                                                    |     |
|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 7.8 | The effect of overlapping extension ( <i>on</i> ) and the number of community assignments ( <i>om</i> ) over the NMI for SLPA and MEMELink. . . . .                                                                                                                                                                                                                                | 137 |
| 7.9 | A comparison between the results achieved by MEMELink and the results presented by Xie et al. for nine distinct methods for overlapping community detection. Figure adapted from Xie, J., Kelley, S., & Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. <i>ACM computing surveys (csur)</i> , 45(4), 43. . . . . | 138 |
| 8.1 | The weighted network formed by the 107 characters and their interactions in the book ‘storm of swords’. . . . .                                                                                                                                                                                                                                                                    | 142 |
| 8.2 | (Left) The solid line shows the degree distribution for the network ‘A storm of swords’, and the dotted blue line shows the trending line following a power law distribution. (Right) The solid line shows the centrality for each node in the network; additionally, the dotted blue line shows the trending line following a power law distribution. . . . .                     | 143 |
| 8.3 | (Left) The count and percentage of characters according to the number of communities. (Centre) The number of nodes for each community. (Right) The chart with the distribution of the number of nodes and the number of community assignments. . . . .                                                                                                                             | 144 |
| 8.4 | The eight overlapping communities detected by MADOC in the ‘Storm of Swords’ network. . . . .                                                                                                                                                                                                                                                                                      | 145 |
| 8.5 | Communities C1 to C8 detected by MADOC highlighted by distinct colours in the ‘Storm of Swords’ network. . . . .                                                                                                                                                                                                                                                                   | 146 |
| 8.6 | (Left) The count and percentage of characters according to the number of communities. (Centre) The number of nodes for each community. (Right) The chart with the distribution of the number of nodes and the number of community assignments. . . . .                                                                                                                             | 147 |
| 8.7 | The nine overlapping communities detected by MEMELink in the ‘Storm of Swords’ network. . . . .                                                                                                                                                                                                                                                                                    | 148 |
| 8.8 | Communities C1 to C9 detected by MEMELink highlighted by distinct colours in the ‘Storm of Swords’ network. . . . .                                                                                                                                                                                                                                                                | 149 |
| 8.9 | (Left) A comparison of the community sizes detected by MEMELink (in green) and MADOC (in blue). (Right) The number of communities per node in the overlapping community structure detected by MEMELink and MADOC. . . . .                                                                                                                                                          | 149 |

8.10 (Left) The overlapping communities of characters in the novel ‘Storm of Swords’ detected by MEMELink. (Right) The overlapping communities of characters in the novel ‘Storm of Swords’ detected by MADOC. . . . . 150

# List of Algorithms

|     |                                                                                                              |     |
|-----|--------------------------------------------------------------------------------------------------------------|-----|
| 3.1 | The general framework of a genetic algorithm. . . . .                                                        | 53  |
| 3.2 | The general framework of a simple local-search based memetic algorithm. . . . .                              | 59  |
| 4.1 | The general framework of MADOC. . . . .                                                                      | 74  |
| 4.2 | A node seed-expansion method to initialise the individuals in MADOC's population. . . . .                    | 79  |
| 4.3 | The label-propagation method used to initialise individuals in MADOC's population. . . . .                   | 80  |
| 4.4 | Modularity recombination crossover used by MADOC. . . . .                                                    | 82  |
| 4.5 | The neighbouring communities mutation used by MADOC. . . . .                                                 | 83  |
| 4.6 | The basic hill-climbing local-search employed by MADOC. . . . .                                              | 86  |
| 4.7 | Deterministic-update annealing for modularity maximisation used by MADOC as a local-search strategy. . . . . | 88  |
| 7.1 | The general framework of MEMELink. . . . .                                                                   | 118 |
| 7.2 | Uniform crossover employed in MEMELink. . . . .                                                              | 123 |
| 7.3 | The neighbouring mutation used by MEMELink. . . . .                                                          | 124 |
| 7.4 | The local-search strategy adopted by MEMELink. . . . .                                                       | 125 |

# ABSTRACT

Intricate relationships exist among the billions of individuals who form our society. The interactions that occur between thousands of genes within our cells comprise our biology. Millions of financial institutions perform billions of transactions daily, creating complex networks of entities. These are just a few examples of the many complex systems surrounding us.

Network science comprehends the collection, management, analysis, interpretation and presentation of complex systems that employ networks. Collections of interconnected items that are usually represented by a graph showing a set of nodes joined by edges.

Complex networks often present community structures where nodes preferentially link to one another. Examples of community structures include groups of friends in society, groups of co-functioning genes in gene networks and groups of similar products in co-purchasing networks, among many others. Detecting the community structure in networks offers important information about the organisation and functioning of such groups. For many phenomena represented by networks, communities can be overlapping, with nodes participating in multiple communities. For example, a person participates in several social organisations; a gene is related to different biological functions; a product can be sold in different markets.

Revealing the community structure in complex networks is no trivial task and can lead to a non-deterministic polynomial-time hardness (NP-hard) computational problem. In this thesis, we approach the overlapping community detection problem using memetic algorithms, metaheuristics that employ a population-based search and local-search inspired by Darwinian principles of natural evolution and Dawkins's notion of a meme defined as a unit of cultural evolution. We detail the construction of two different memetic algorithms, present computational results, compare our methods with other state-of-the-art metaheuristics and present applications of our methods as case studies.