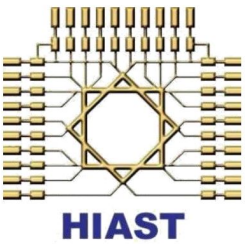
Syrian Arab Republic
Higher Institute for Applied Sciences and Technology
Informatics Department
Fourth Year



Using Genetic Algorithms to Find Approximations for the Minimum Vertex Cover Problem

Keywords: minimum vertex cover, NP-complete, genetic algorithms, combinatorial optimization, approximation algorithms.

Author: *Farouk Hjabo*
Academic Supervisor: *Dr. Said Desouki*
General Supervisor: *Dr. Kadan Aljoumaa*
Langauge Supervisor: *Mr. Fahmi Alammareen*

March 3, 2018

Abstract

Remarkable results obtained from the application of a genetic algorithm to the minimum vertex cover problem are reported, and tested against the results of the best known heuristic algorithm for graphs of sizes 100 and 200 nodes. In contrast to many other approximation algorithms that use problem-specific knowledge, the approach presented relies on a graded penalty term component of the fitness function to penalize infeasible solutions.

"It is not the strongest of the species that survives, nor the most intelligent, but the one most responsive to change."

— Charles Darwin

Contents

List of Figures

List of Tables

Abbreviations

GA	**G**enetic **A**lgorithm
mvc	**M**inimum **V**ertex **C**over
mvcp	**M**inimum **V**ertex **C**over **P**roblem

1 Introduction

Once the NP-hardness of a combinatorial optimization problem is established, the search for an optimal solution is abandoned. Our goal then becomes one of finding good heuristics for that problem. That is, we try to find approximation algorithms that have polynomial time complexities and return near-optimal solutions. In most cases, the heuristic is problem dependent; it is tailored to specific problem it is trying to solve, and it is tightly coupled with observations made about the problem. The *minimum vertex cover* problem is a typical example of such problems [19]. Due to its numerous applications especially in various matching problems, the problem is not abandoned and it has many approximation algorithms.

In this work, we present an alternative approach that uses *genetic algorithms* as a generalized heuristic for solving NP-hard combinatorial optimization problems. These algorithms have been successfully applied to a broad range of problems. This wide range can be tackled by genetic algorithms mainly due to the fact that these algorithms work with an encoding of the problem rather than the characteristics of that problem. Remarkable results were obtained and compared to the results of the best known heuristic algorithm for the mvcp.

The outline of this paper is as follows: Firstly, section 2 gives a short introduction to genetic algorithms, and explains briefly how they work. Section 3 introduces the mvcp with two of its main heuristics, and explains its representation for an application of the genetic algorithm. Lastly, section 4 presents our experimental results and points out the notable observations that can be drawn from these results.

2 Genetic Algorithms

2.1 The Intuition Behind GAs

Genetic Algorithms (GAs) [3, 8] are population based search algorithms, where by repeated use of genetic operations, such as ***mutation***, ***selection***, ***crossover***, etc... Successive new generations of better populations in the direction of search objectives are created. They are inspired by Darwinian principles based on natural evolution. In other words, the main idea behind genetic algorithms is that only the ***fittest*** individuals will survive among generations of evolution.

Main advantage of genetic algorithms is that they ideally do not make any assumption about the underlying problem, hence they are suitable to tackle a wide

range of diverse problems in engineering, art, biology, economics, marketing, genetics, operations research, robotics, social sciences, physics, politics, chemistry, etc... [3, ch. 6]

So what is a GA? A typical GA consists of the following [3, ch. 1]:

1. a number, or ***population***, of candidate solutions to the problem.
2. a way of calculating how good or bad the individual solutions within the population are. i.e. how ***fit*** an individual is?
3. a method for mixing fragments of the better solutions to form new, on average even better solutions. Which permits the population to ***evolve*** naturally.

2.2 Typical GA Schema

The basic iteration cycle of a genetic algorithm proceeds on a population of individuals, each of which represents a search point in the space of potential solutions of a given optimization problem. In case of a canonical genetic algorithm, each individual is a binary vector $\vec{x} = (x_1, \ldots, x_n) \in \{0,1\}^n$ of fixed length n. The fitness function $f : \{0,1\}^n \to \mathbb{R}$ provides a quality measure which is used by the selection procedure to direct the search towards regions of the search space where the average fitness of the population increases. The recombination operator allows for the exchange of information between different individuals, and mutation introduces innovation into the population. This cycle is repeated until a termination criterion is fulfilled. In most cases, the algorithm is terminated after a certain number of iterations of the basic cycle or when a satisfactory fitness level has been reached.
Refer to figure 1 for a visual representation of this process, and refer to algorithm 1 for a more concise description.

Initialization The population size (denoted as u) depends on the nature of the problem, but typically contains several hundreds or thousands of possible solutions. Often, the initial population is generated randomly, allowing the entire range of possible solutions (the search space). Occasionally, the solutions may be 'seeded' in areas where optimal solutions are likely to be found.

Selection Normally, selection in genetic algorithms is a probabilistic operator which uses the relative fitness $p_s(\vec{x_i}) = f(\vec{x_i}) / \sum_{j=1}^{u} f(\vec{x_j})$ to serve as selection probabilities (u denotes the population size). This selection operator is called proportional selection. It is commonly used; as intuitively the more fit a solution is the more we want it to affect the next generations.

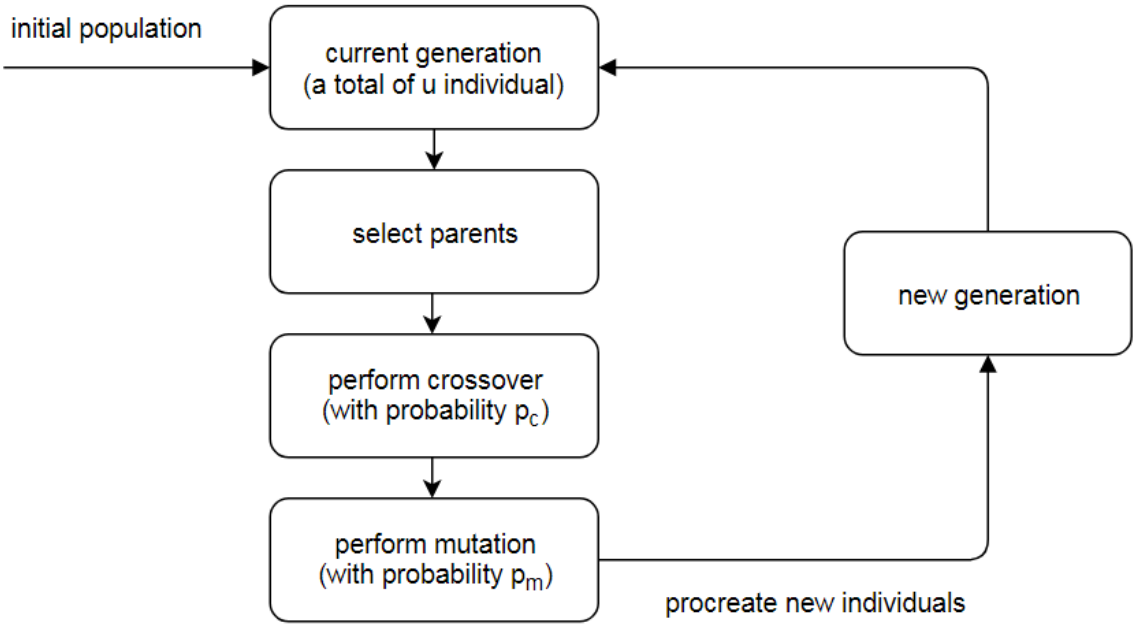


Figure 1: General schema for genetic algorithms

If the problem under consideration is a minimization one, or if the fitness function can take negative values, then $f(\vec{x_i})$ has to be linearly transformed before calculating selection probabilities. This technique known as linear dynamic scaling is commonly used in genetic algorithms [8, pp. 123–124].

Crossover The recombination (crossover) operator allows for the exchange of information between different individuals. The original one-point crossover [10] works on two parent individuals by choosing a crossover point $\chi \in \{1, \ldots, n-1\}$ at random and exchanging all bits after the χ^{th} one between both individuals. Therefore, two new off-springs will be formed.
For example, if the two parents were encoded as $(111, 000)$ and we chose $\chi = 2$ then the new two off-springs will be $(110, 001)$. The crossover rate p_c (e.g. $p_c \approx 0.6$) determines the probability to undergo crossover (see algorithm 1 line 9), this allows the two parents to be present in the next generation; as crossover is not mandatory. Moreover, the value of p_c indicates the ratio of new off-springs in the next generation. The crossover operator can be extended to a generalized multi-point crossover or even to uniform crossover, where it is randomly decided for each bit whether to exchange it or not. The strong mixing effect introduced by uniform crossover is sometimes helpful to overcome local optima.

Mutation Innovation (new information) is introduced into the population by means of mutation, which works by inverting bits with a small probability p_m (e.g. $p_m \approx 0.001$) (see algorithm 1 lines 10 11 for clarity). Though mutation is often interpreted as a rather unimportant operator in genetic algorithms [10], recent theoretical work gives strong evidence for an appropriate choice of a mutation rate $p_m = 1/n$ on many problems [2, 13].

Algorithm 1 GenticAlgorithm(n, u, p_c, p_m, f)

```
 1: P ← {}                                          ▷ P contains the population
 2: loop u times                                                ▷ initialization
 3:     add a randomly generated binary vector of length n to P
 4: end loop
 5: while termination criterion is not fulfilled do
 6:     Q ← {}                                      ▷ Q will hold the new generation
 7:     loop u/2 times
 8:         select x, y from P using proportional selection          ▷ selection
 9:         with probability p_c perform crossover between x, y      ▷ crossover
10:         for each bit in x flip it with probability p_m           ▷ mutation
11:         for each bit in y flip it with probability p_m           ▷ mutation
12:         add x, y to Q
13:     end loop
14:     P ← Q
15: end while
```

3 The Minimum Vertex Cover Problem

The problem of finding a minimum vertex cover is a classical optimization problem in computer science and is a typical example of an NP-complete optimization problem [19] that has approximation algorithms. The mvcp of an undirected graph $G = (V, E)$ where V is the set of vertices and E denotes the set of edges, consists in finding the smallest subset $V' \subseteq V$ such that $\forall \langle i, j \rangle \in E$, we have $i \in V'$ or $j \in V'$ (or both). V' is said to be a vertex cover of G.

The following is a formal definition of the mvcp in which we make use of Stinson's terminology for combinatorial optimization problems [18]:

Problem instance: A graph $G = (V, E)$, where $V = \{1, 2, \ldots, n\}$ is the set of vertices and $E \subseteq V \times V$ the set of edges. An edge between vertices i, j is denoted by

the pair $\langle i, j\rangle \in E$. We define the adjacency matrix e_{ij} according to:

$$e_{ij} = \begin{cases} 1 & \text{if } \langle i, j\rangle \in E \\ 0 & \text{otherwise} \end{cases}$$

On particular, we have $e_{ii} = 0$.

Feasible solution: A set $V' \subseteq V$ such that $\forall\langle i, j\rangle \in E, i \in V' \vee j \in V'$

Objective function: The size $|V'|$ of the vertex cover V'.

Optimal solution: a vertex cover V' that minimizes $|V'|$.

An interesting property of the mvcp is that by solving it, we also have a solution for two other graph problems: the *maximum independent set problem* (given $G = (V, E)$, find the largest subset $V' \subseteq V$ such that $\forall i, j \in V', \langle i, j\rangle \notin E$) and the *maximum clique problem* (given $G = (V, E)$, find the largest subset $V' \subseteq V$ such that $\forall i, j \in V', \langle i, j\rangle \in E$) The close relationship between these problems is characterized by the following lemma [7].

Lemma 1

For any graph $G = (V, E)$ and $V' \subseteq V$, the following statements are equivalent:

- *V' is the minimum vertex cover of G.*
- *$V - V'$ is the maximum independent set in G.*
- *$V - V'$ is the maximum clique in $\bar{G} = (V, \bar{E})$, where $\bar{E} = \{\langle i, j\rangle \in V \times V : \langle i, j\rangle \notin E\}$*

Consequently, one can obtain a solution of the maximum independent set problem by taking the complement of the solution to the minimum vertex cover problem. A solution to the maximum clique problem is obtained by taking the complement of the minimum vertex cover of $\bar{G}$.

3.1 Related Algorithms

Since we are interested in covering all the edges of the given graph by using as few nodes as possible, one might be tempted to use a greedy-based heuristic to tackle the

mvcp. The algorithm consists in repeatedly selecting a vertex of highest degree (the node that covers as many of the remaining edges as possible), and removing all of its incident edges. This is not a good strategy as was demonstrated by Papadimitriou and Steiglitz [15, p. 407].

They considered regular graphs*, each of which consists of three levels. The first two levels have the same number of nodes while the third level has two nodes less than the number of nodes found on the previous two levels. More precisely, each graph consists of $n = 3k + 4$ ($k \geq 1$) nodes; $k + 2$ nodes on the first level, labeled $1, 2, \ldots, k + 2$; followed by $k + 2$ nodes on the second level, labeled $k + 3, \ldots, 2k + 4$; while the k nodes of the third level are labeled $2k + 5, \ldots, 3k + 4$. The regular graph for $k = 3$ can be found in figure 2. A minimum vertex cover is obtained by choosing all the nodes of the second level, but the greedy strategy would start with the nodes of the third level, since these have highest degree. Consequently, the greedy strategy finds a solution of size $2k + 2$ (since it has to select $k + 2$ nodes in addition to the nodes of the third level), while the optimal solution has size $k + 2$.

However this algorithm is not totally useless. It may be shown that it always achieves the ratio $\Theta(\lg n)^\dagger$, where n is the number of vertices. That is, if $C^*(n)$ is the size of the optimal solution and $C(n)$ is the size of the solution found by the greedy algorithm. Then we have $\frac{C(n)}{C^*(n)} = \Theta(\lg n)$ [14, p. 323].

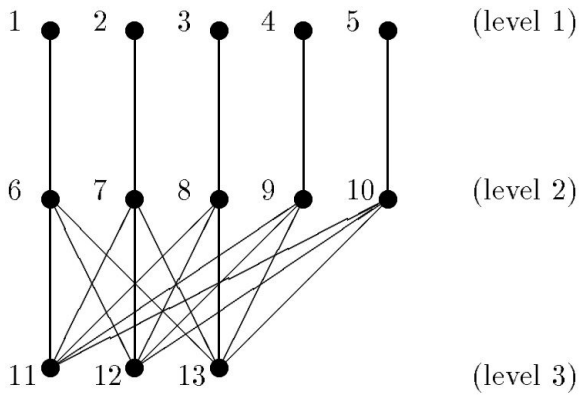


Figure 2: Construction of the regular graph after Papadimitriou and Steiglitz with $k = 3$

Fortunately, a simpler and better algorithm exists. The following simple algorithm 2 is surprisingly the best approximation algorithm known for the mvcp [14, p. 301].

This is a *2-approximation algorithm*. Which means that the size of the vertex cover it returns is guaranteed to be no more than twice the size of an optimal vertex

*by "regular graphs" we don't refer to the definition where each vertex has the same degree.

$^\dagger f(n) = \Theta(g(n)) \iff \exists \alpha, \beta, n_0 : \forall n \geq n_0,\ \alpha \cdot g(n) \leq f(n) \leq \beta \cdot g(n)$ [4, pp. 41–43]

Algorithm 2 vercov($G = (V, E)$)

```
1: C ← {}                                   ▷ C contains the vertex cover being constructed
2: E' ← E
3: while E' ≠ φ do
4:     randomly choose ⟨u, v⟩ ∈ E'
5:     C ← C ∪ {u, v}
6:     E' ← E' − {⟨x, y⟩ ∈ E' : x = u ∨ y = v ∨ x = v ∨ y = u}
7:                                  ▷ remove from E' every edge incident on either u or v
8: end while
9: return C
```

cover [4, p. 968]. More precisely, if C^* is the size of the optimal solution and C is the size of the solution returned by algorithm 2 then the inequality $C \leq 2C^*$ must hold (notice that here the approximation doesn't depend on the number of vertices).

3.2 GA for the mvcp

In order to apply genetic algorithm to solve the mvcp, a cover V' is represented by a binary vector $\vec{x} = x_1 x_2 \dots x_n$ where $x_i = 1$ if the i^{th} node is in V', and $x_i = 0$ if it is not. Using this representation, genetic algorithm is applied to minimize the following fitness function [11]:

$$\begin{aligned} f(\vec{x}) &= \sum_{i=1}^{n} \left(x_i + n \cdot (1 - x_i) \cdot \sum_{j=i}^{n} (1 - x_j) e_{ij} \right) \\ &= \sum_{i=1}^{n} x_i + n \cdot \sum_{i=1}^{n} \sum_{j=i}^{n} (1 - x_i) \cdot (1 - x_j) \cdot e_{ij} \end{aligned} \tag{1}$$

The term $\sum_{i=1}^{n} x_i$ of $f(\vec{x})$ determines the size of the potential vertex cover represented by $\vec{x}$, while the term $n \cdot \sum_{i=1}^{n} \sum_{j=i}^{n} (1 - x_i) \cdot (1 - x_j) \cdot e_{ij}$ penalizes sets V' that are not covers by adding a penalty of magnitude n for each edge e_{ij} for which $i \notin V' \wedge j \notin V'$ (note that e_{ii} is previously defined to be zero). Consequently, any infeasible solution will have a fitness greater than or equal to n. And for any feasible solution the second term will drop to zero, making the fitness value of the worst feasible solution at most n.

This fitness function was developed according to the following design principles that are important for a successful penalty function approach [12, 16]:

- Infeasible binary vectors are guaranteed to yield fitness values which are inferior to fitness values of the worst feasible solutions.

- The penalty should be graded, i.e., fitness values should improve as solutions approach feasible regions of the search space. In other words, the more edges a solution covers, the fitter it is. notice that this is guaranteed by the second term in equation 1, where for each uncovered edge the fitness is penalized with the value of n.

The experiments reported in section 4.1 are performed by using a genetic algorithm with a population size of $u = 50$, mutation rate $p_m = 1/n$, crossover rate $p_c = 0.6$, proportional selection, and two-point crossover [11]. As reported in [6, 17], the latter is expected to perform better than the traditional one-point crossover. A bound for the total number of fitness function evaluations is set as the termination criterion [11], see section 4.1 for details.

4 Experiments

For the experimental tests, the genetic algorithm software package GENEsYs is used [1]. This implementation is based on the widely used GENESIS software by Grefenstette [9], but allows for more flexibility concerning genetic operators and data monitoring. The parameter settings for the experiments are given in section 3.2.

4.1 Experimental Results

Test set 1: Multiple problem instances are used to demonstrate the applicability of GAs to tackle the mvcp. The results in table 1 and table 2 correspond to the test set 1, which consists of five random graphs of size $n = 100$ with different edge densities d. Densities $d \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ are used respectively. An edge density of $d = 0.1$ means that an edge is placed between two nodes with a probability of 0.1. Moreover, we construct the graphs in a way that guarantees an optimum of quality 55 [11]. i.e. we use algorithm 3 with $k = 55$ to generate random graphs.
In addition, the regular graphs introduced by Papadimitriou and Steiglitz and described in section 3.1 are used. Recall that these graphs contain $n = 3k + 4$ ($k \geq 1$) nodes distributed on three levels. They can be scaled up by choosing high values for k. A graph instance of the regular graph of size $n = 100$ ($k = 32$) is chosen.

Termination: As a termination criterion for the genetic algorithm, a bound for the total number of fitness function evaluations is set. For example The total number of function evaluations per single run is chosen to be $2 \cdot 10^4$ in table 1. That is, the

Algorithm 3 GenerateRandomGraph(n, d, k)

```
1: randomly select V' = {i_1, ..., i_k} ⊆ V = {1, ..., n}
2:         ▷ randomly select k different nodes, these nodes will hold a potential mvc
3: for i = 1 to n − 1 do
4:     for j = i + 1 to n do
5:         if (Random(0, 1) ≤ d) ∧ (i ∈ V' ∨ j ∈ V') then
6:             e_ij = 1
7:         else
8:             e_ij = 0
```

algorithm terminates when $2 \cdot 10^4$ evaluations of the fitness function had occurred. This bound is indicated as an index t in the notation $f_t(\vec{x})$. Consequently, only a small fraction of the search space is explored by the genetic algorithm. This fraction is about $2{\cdot}10^4/2^{100} \approx 1.6{\cdot}10^{-24}\%$ for test set 1. And it is about $4{\cdot}10^4/2^{200} \approx 2.5{\cdot}10^{-54}\%$ for test set 2.

Table 1: For each graph in test set 1, a total of $N = 100$ independent runs of the genetic algorithm is performed. The results are summarized in table 1 for the best fitness values that were encountered during the 100 runs. For each problem instance, the different fitness values that were obtained and their frequencies are recorded. Furthermore, the average fitness value $\bar{f}$ over all 100 runs is indicated at the bottom of the table.
For example, the first column in table 1 indicates that the best obtained fitness value in 1 of the 100 runs was $f(\vec{x}) = 53$, the best obtained fitness value in 1 of the 100 runs was $f(\vec{x}) = 54$, the best obtained fitness value in 3 of the 100 runs was $f(\vec{x}) = 55$, and so on... And the average of these values is indicated by $\bar{f}$, which is the average value of the best obtained fitness values over the 100 runs.

Table 2: For each of the problems in the test set 1, a total of $N = 100$ independent runs of the vercov heuristic is performed. The results are summarized in table 2. Notice that for one problem instance, the results may differ in each run due to the non-determinism of the algorithm (see algorithm 2 line 4).

Test set 2: To test the behaviour of the genetic algorithm as well as the vercov heuristic for an even larger problem size, The same experiments were performed for larger graphs. The results in table 3 and table 4 correspond to test set 2, which consists of five random graphs of size $n = 200$ with different edge densities d. Densities $d \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ are used. Moreover, we construct the graphs in a way that

mvcp100-01		mvcp100-02		mvcp100-03		mvcp100-04		mvcp100-05		PS100	
$f_{2\cdot 10^4}(\vec{x})$	N	$f_{2\cdot 10^4}(\vec{x})$	N	$f_{2\cdot 10^4}(\vec{x})$	N	$f_{2\cdot 10^4}(\vec{x})$	N	$f_{2\cdot 10^4}(\vec{x})$	N	$f_{2\cdot 10^4}(\vec{x})$	N
53	1	55	34	55	77	55	96	55	99	34	65
54	1	57	3	56	1	67	1	83	1	66	35
55	3	59	2	59	6	68	1				
56	6	60	3	63	3	75	1				
57	4	61	10	64	3	90	1				
58	4	62	1	66	1						
59	9	63	8	67	1						
60	4	64	1	68	1						
61	6	65	1	70	1						
62	12	66	6	71	1						
63	5	67	6	74	1						
> 63	45	> 67	25	> 74	4						
$\bar{f} = 62.61$		$\bar{f} = 62.75$		$\bar{f} = 57.62$		$\bar{f} = 55.80$		$\bar{f} = 55.28$		$\bar{f} = 45.20$	

Table 1: Experimental results obtained by the genetic algorithm for five random graphs of size $n = 100$ with edge density: $d = 0.1$ ("mvcp100-01"), $d = 0.2$ ("mvcp100-02"), $d = 0.3$ ("mvcp100-03"), $d = 0.5$ ("mvcp100-04"), $d = 0.5$ ("mvcp100-05") and the regular graph of size $n = 100$ ($k = 32$) from Papadimitriou and Steiglitz ("PS100").

mvcp100-01		mvcp100-02		mvcp100-03		mvcp100-04		mvcp100-05		PS100	
$f(\vec{x})$	N	$f(\vec{x})$	N	$f(\vec{x})$	N	$f(\vec{x})$	N	$f(\vec{x})$	N	$f(\vec{x})$	N
80	6	80		80		80	1	80		66	100
82	17	82	1	82		82		82	1		
84	23	84	2	84	1	84	1	84			
86	20	86	4	86	6	86		86	5		
88	17	88	18	88	12	88	5	88	7		
90	14	90	17	90	15	90	13	90	10		
92	1	92	31	92	18	92	24	92	21		
94	2	94	20	94	28	94	34	94	27		
96		96	4	96	16	96	19	96	17		
98		98	1	98	4	98	3	98	9		
100		100		100		100		100	3		
$\bar{f} = 86.46$		$\bar{f} = 89.22$		$\bar{f} = 92.22$		$\bar{f} = 92.96$		$\bar{f} = 93.12$			

Table 2: Experimental results obtained by the vercov heuristic for five random graphs of size $n = 100$ with edge density: $d = 0.1$ ("mvcp100-01"), $d = 0.2$ ("mvcp100-02"), $d = 0.3$ ("mvcp100-03"), $d = 0.5$ ("mvcp100-04"), $d = 0.5$ ("mvcp100-05") and the regular graph of size $n = 100$ ($k = 32$) from Papadimitriou and Steiglitz ("PS100").

guarantees an optimum of quality 110 (see algorithm 3, we set $k = 110$). In addition to a graph instance of the regular graphs introduced by Papadimitriou and Steiglitzof of size $n = 220$ ($k = 66$). In this case, the genetic algorithm was allowed to run for $4 \cdot 10^4$ function evaluations.
The corresponding results obtained by the genetic algorithm are shown in table 3, while the results from the vercov heuristic are shown in table 4. See the previous paragraphs for a better understanding of these results.

mvcp200-01		mvcp200-02		mvcp200-03		mvcp200-04		mvcp200-05		PS202	
$f_{4\cdot10^4}(\vec{x})$	N	$f_{4\cdot10^4}(\vec{x})$	N	$f_{4\cdot10^4}(\vec{x})$	N	$f_{4\cdot10^4}(\vec{x})$	N	$f_{4\cdot10^4}(\vec{x})$	N	$f_{4\cdot10^4}(\vec{x})$	N
110	2	110	55	110	95	110	99	110	100	68	60
113	1	120	10	128	1	140	1			134	40
116	4	121	7	129	6						
117	2	122	2	130	2						
119	3	123	7	134	1						
120	1	125	1								
121	2	126	1								
122	3	127	1								
124	3	129	1								
125	6	132	2								
126	2	134	1								
> 126	71	> 134	12								
$\bar{f} = 132.50$		$\bar{f} = 119.07$		$\bar{f} = 111.01$		$\bar{f} = 110.30$		$\bar{f} = 110.00$		$\bar{f} = 108.00$	

Table 3: Experimental results obtained by the genetic algorithm for five random graphs of size $n = 200$ with edge density: $d = 0.1$ ("mvcp200-01"), $d = 0.2$ ("mvcp200-02"), $d = 0.3$ ("mvcp200-03"), $d = 0.5$ ("mvcp200-04"), $d = 0.5$ ("mvcp200-05") and the regular graph of size $n = 202$ ($k = 66$) from Papadimitriou and Steiglitz ("PS202").

4.2 Experimental Conclusions

From these tables, a number of interesting conclusions can be drawn. For the random graphs used to compare the genetic algorithm and the vercov heuristic, it is known by construction that an optimum of quality 55 exists for test set 1, (respectively, an optimum of quality 110 exists for the test set 2). This optimum is likely to be the global optimum. And it is found by the genetic algorithm at least once within the $N = 100$ runs that were performed.

Moreover, the randomly constructed problems quickly become simpler for the genetic algorithm when the edge density is increased. For an edge density of $d \in \{0.4, 0.5\}$, the genetic algorithm almost surely finds the global optimum of the problems, while vercov heuristic gives a relatively terrible results.

mvcp200-01		mvcp200-02		mvcp200-03		mvcp200-04		mvcp200-05		PS202	
$f(\vec{x})$	N	$f(\vec{x})$	N	$f(\vec{x})$	N	$f(\vec{x})$	N	$f(\vec{x})$	N	$f(\vec{x})$	N
172	2	172		172		172		172		134	100
174	3	174		174	1	174		174			
176	13	176	1	176	1	176	1	176	1		
178	14	178	1	178	3	178	4	178	1		
180	23	180	5	180	5	180	1	180	5		
182	14	182	11	182	5	182	9	182	5		
184	15	184	18	184	17	184	8	184	5		
186	7	186	16	186	19	186	16	186	13		
188	5	188	24	188	16	188	16	188	11		
190	3	190	15	190	16	190	16	190	21		
192	1	192	4	192	9	192	15	192	19		
> 192	0	> 192	5	> 192	8	> 192	14	> 192	19		
$\bar{f} = 180.98$		$\bar{f} = 186.46$		$\bar{f} = 186.92$		$\bar{f} = 188.06$		$\bar{f} = 189.16$			

Table 4: Experimental results obtained by the vercov heuristic for five random graphs of size $n = 200$ with edge density: $d = 0.1$ ("mvcp200-01"), $d = 0.2$ ("mvcp200-02"), $d = 0.3$ ("mvcp200-03"), $d = 0.5$ ("mvcp200-04"), $d = 0.5$ ("mvcp200-05") and the regular graph of size $n = 202$ ($k = 66$) from Papadimitriou and Steiglitz ("PS202").

In case of the regular graph after Papadimitriou and Steiglitz, the genetic algorithm is able to find the global optimum in about 2/3 of all runs, while the remaining runs identify a solution of quality $2k + 2$. Whereas, vercov heuristic gives a solution of quality $2k + 2$ in all of the 100 runs as expected [15]. Recall that the graphs are constructed so as to force the vercov heuristic into yielding solutions of quality $2k+2$. The random choice of edges (line 4 of algorithm 2) implies that nodes from either layers one and two, or layers two and three are involved in the solution being constructed, which in turn implies that two layers have to be part of the solution found by the algorithm.

For the randomly constructed graphs, the vercov heuristic performs poorly. It barely finds solutions of quality 80 (for graphs of size 100) and 172 (for graphs of size 200). Moreover, notice that the quality of its solutions are dominated by the quality of the solutions found by the genetic algorithm; the worst solutions found by the genetic algorithm are almost as good as the best solutions found by the vercov heuristic.

The average performance of the vercov heuristic (surprisingly) decreases as the edge density is increased; i.e. the problems become even harder for the vercov heuristic while they become much simpler for the genetic algorithm when the edge density

increases. The average fitness found by the genetic algorithm is notably better than the one found by the vercov heuristic. One can demonstrate this by calculating the ratio of the absolute (or equivalently, the relative) error. For example, for the first problem instance, let $\bar{f}_1, \bar{f}_2$ be the average found by the genetic algorithm, vercov heuristic respectively. We have $(\bar{f}_2 - 55)/(\bar{f}_1 - 55) = (86.46 - 55)/(62.61 - 55) \approx 4 = 400\%$ that is, the error was reduced by a factor of 4. This is further visualized in figure 3 for the random graphs in test set 1 (figure on the left) and for the random graphs in test set 2 (figure on the right), where the average value found by each algorithm for each graph instance is plotted.

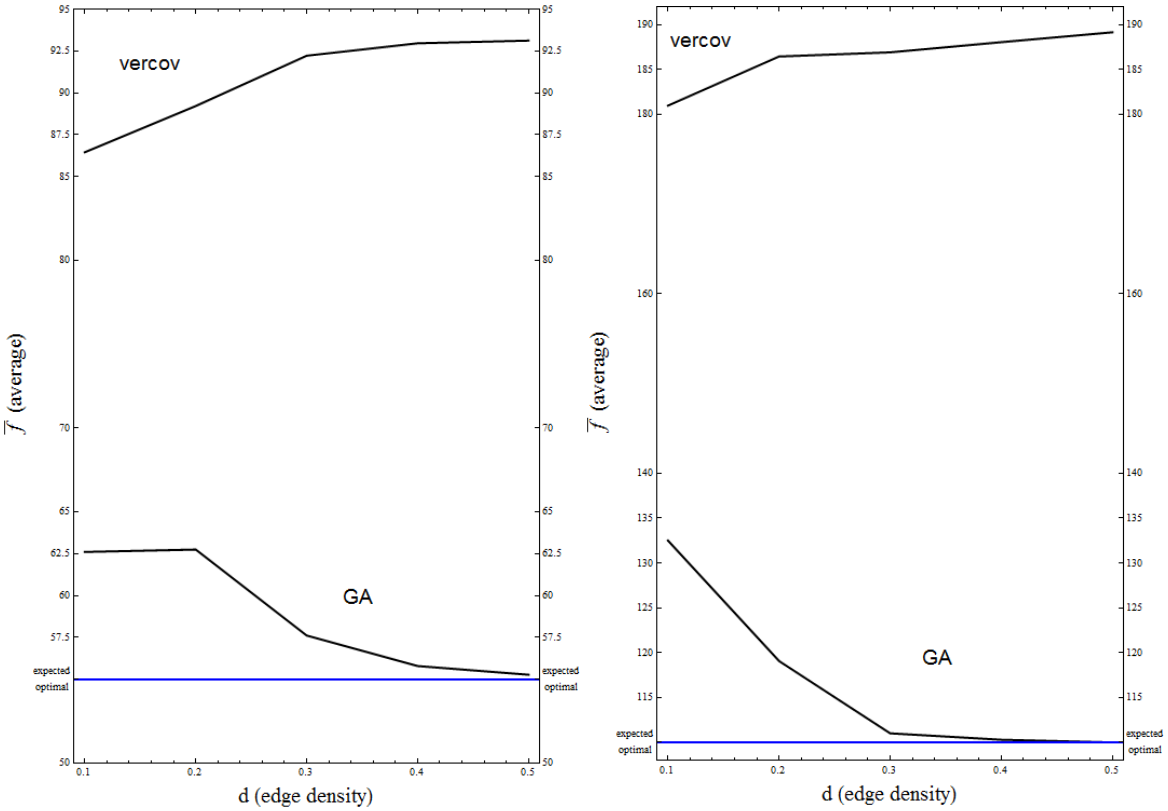


Figure 3: Average fitness value found by GA and vercov heuristic plotted against the edge density for each random graph in test set 1 (the left figure) and for each random graph in test set 2 (the right figure).

5 Conclusion

This work has demonstrated that genetic algorithms can be used in a fairly straightforward way to find good approximative solutions for NP-hard problems, namely the minimum vertex cover problem. The robustness of the presented solution along with

the positive results obtained support the idea that genetic algorithms are a desirable approach for tackling these type of problems. To tackle a new problem, rather than having to construct tailored heuristics to handle the problem under consideration, we advocate the use of genetic algorithms where the only change to perform is the formulation of a new fitness function.

References

[1] T. Back. "GENEsYs 1.0. Software distribution". In: *Systems Analysis Research Group, LSXI, University of Dortmund, Germany* (1992).

[2] T. Bäck. "The interaction of mutation rate, selection, and self-adaptation within a genetic algorithm". In: *Proc. 2nd Conference of Parallel Problem Solving from Nature, 1992*. Elsevier Science Publishers. 1992.

[3] D. A. Coley. *An Introduction to Genetic Algorithms for Scientists and Engineers*. World Scientific, 1999.

[4] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction To Algorithms, 2nd Edition*. MIT electrical engineering and computer science series. MIT Press, 2001.

[5] L. Davis. *Handbook of genetic algorithms*. VNR computer library. Van Nostrand Reinhold, 1991.

[6] L. J. Eshelman, R. A. Caruana, and J. D. Schaffer. "Biases in the crossover landscape". In: *Proceedings of the third international conference on Genetic algorithms*. Morgan Kaufmann Publishers Inc. 1989, pp. 10–19.

[7] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. Books in mathematical series. W. H. Freeman, 1979.

[8] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Artificial Intelligence. Addison-Wesley Publishing Company, 1989.

[9] J. Grefenstette. "GENESIS: A system for using genetic search procedures". In: *Proceedings of a Conference on Intelligent Systems and Machines*. 1984, pp. 161–165.

[10] J. H. Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

[11] S. Khuri and T. Bäck. "An evolutionary heuristic for the minimum vertex cover problem". In: *Genetic Algorithms within the Framework of Evolutionary Computation–Proc. of the KI-94 Workshop*. Available at: `https://www.researchgate.net/publication/2581134_An_Evolutionary_Heuristic_for_the_Minimum_Vertex_Cover_Problem`. Saarbrücken, Germany. 1994, pp. 86–90.

[12] S. Khuri, T. Bäck, and J. Heitkötter. "An Evolutionary Approach to Combinatorial Optimization Problems." In: *ACM Conference on Computer Science*. Citeseer. 1994, pp. 66–73.

[13] H. Muhlenbein. "How genetic algorithms really work: I. mutation and hillclimbing". In: *Proc. 2nd Int. Conf. on Parallel Problem Solving from Nature, 1992*. Elsevier. 1992.

[14] C. H. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1994.

[15] C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Dover Books on Computer Science. Prentice Hall, 1982.

[16] J. T. Richardson, M. R. Palmer, G. E. Liepins, and M. R. Hilliard. "Some guidelines for genetic algorithms with penalty functions". In: *Proceedings of the 3rd international conference on genetic algorithms*. Morgan Kaufmann Publishers Inc. 1989, pp. 191–197.

[17] J. D. Schaffer. "A study of control parameters affecting online performance of genetic algorithms for function optimization". In: *San Meteo, California* (1989).

[18] D. R. Stinson. *An Introduction to the Design and Analysis of Algorithms*. Charles Babbage Research Centre, 1985.

[19] E. W. Weisstein. *"Minimum Vertex Cover"*. From MathWorld–A Wolfram Web Resource. URL: `http://mathworld.wolfram.com/MinimumVertexCover.html` (visited on 02/26/2018).