Noname manuscript No.
(will be inserted by the editor)

Quality certification of vertex cover heuristics on real-world networks

Fabrice Lécuyer

Received: date / Accepted: date

Keywords Complex networks · Vertex cover · Quality certification · Heuristics

1 Addressing the vertex cover problem on real-world networks

The vertex cover problem consists in finding a subset of nodes that touch all the edges of a graph. Numerous real-world applications are equivalent to finding a vertex cover with minimum number of nodes: vaccinating as few people as possible to prevent any transmission of a virus, destroying as few routers as possible to shutdown a network, using as little energy as possible to provide wireless connectivity to a given area... Yet, finding the exact result is generally not feasible under time constraints: no polynomial algorithm is known [1], even when the graph is planar or when nodes have small degree.

When facing this problem on datasets that represent large complex networks, engineers have two options. The first one consists in using a preprecessing method that reduces the graph with specific rules before applying an exact exponential-time algorithm. In 2019, a programming challenge fostered efforts in this direction and rewarded the implementation of [3]. If the network has suitable properties, it can lead to an exact minimum cover in reasonable time; but in general, there is no guarantee that the execution terminates in the next hours, days or years.

The alternative option is to use fast and intuitive algorithms that have no theoretical guarantees, called heuristics. Several of them are combined in the linear-time implementation of [4]. Their execution time can be predicted and the quality of their result can be excellent, but it is not mathematically guaranteed: there is no indication on how far the heuristic result is from optimum.

This work is funded by the French National Agency of Research through ANR FiT LabCom.

Fabrice Lécuyer, E-mail: fabrice.lecuyer@lip6.fr
Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

To bridge the gap, we propose a method that certifies the quality of a heuristic on a given network. The quality certification takes a network and gives both an approximate result and a certificate of its quality, defined as the ratio between the heuristic result and a bound on the optimum value. For example, the shortest path between two cities is lower-bounded by the distance as the crow flies; the certified quality of a path is then given by the ratio between its length and the lower-bound. To obtain lower-bounds for vertex cover, we design the two strategies presented in Section 2. The first one uses the well-known dual problem of maximum matching. The second uses a greedy clique partition that particularly fits complex networks.

The experiments of Section 3 test the method on 114 real-world networks with up to three billion edges: we certify that the results of state-of-the-art heuristics for the minimum vertex cover problem are within 1% of the optimum value on two thirds of the networks. This work shows that valuable quality certificates can be given for existing heuristics on specific networks without loosing on scalability: both the heuristic and the certification take linear time. It outlines the best practice of providing certifications for heuristics in general. As it may generalise to other algorithmic problems, it opens a door for further research and for deployment in real-world applications.

2 Designing lower-bounds to certify the quality of a vertex cover

Given a graph of n nodes, finding the size c^* of a minimum vertex cover is NP-hard. Various heuristics can be used to obtain a *small* but not *minimum* cover of size $c \geq c^*$. Besides, vertex cover has a well-known 2-approximation algorithm that implies a lower-bound on c^*: for any maximal set of e independent edges, a vertex cover needs at least one node to cover each of these e edges, and at most two nodes to cover all the graph. This gives: $e \leq c^* \leq 2e$.

To obtain a useful lower-bound, the aim is to compute the maximum value of e^*, also called the maximum matching number. This problem can be solved in polynomial-time by the blossom algorithm [5]. Faster heuristics can find a close but smaller value e, in particular a linear greedy algorithm: as long as independent edges are in the graph, it selects the one that has the less neighbouring edges.

Definition 1 (Certification by matching) Given a vertex cover of c nodes and a matching of e edges, the quality certification method guarantees that the vertex cover is within factor μ of the minimum cover, with the quality ratio μ defined as:

$$\mu = \frac{c}{e} \qquad \text{then} \qquad c \leq \mu c^*$$

However, as we will see in Section 3, this method does not succeed on denser graphs. In particular, when there is a clique of $2k$ nodes, a matching can have at most k edges while a vertex cover has at least $2k-1$ nodes: for high k, the quality ratio tends towards 2, which is not better than the mathematical

guarantee in any graph. Still, the certification by matching is an important building block for the next bounding heuristic that we propose below.

With their strong community structure, complex networks are known to contain high numbers of cliques [2]. From this observation, we propose a new lower-bounding technique based on a partition of the graph into x independent cliques. To cover all the edges of a clique, a vertex cover needs to contain at least all but one of the nodes of the clique. Summing over the x cliques, we obtain $c^* \geq n - x$.

Finding a small value x^* is called the clique cover problem and is NP-hard; but a greedy linear method exists: it grows a clique progressively by selecting adjacent nodes of small degree. When the clique cannot grow, it is added to the partition. Note that a matching is a particular instance of a clique partition: it partitions the nodes into independent edges (cliques of two nodes) and isolated nodes (cliques of one node). Thus, we can always obtain x such that $n - x \geq e$. In the end, we have the following inequalities:

$$e \leq e^* \leq n - x \leq n - x^* \leq c^* \leq c \leq 2e$$

Definition 2 (Certification by cliques) Given a vertex cover of c nodes and a partition of the nodes into x cliques, the quality certification method guarantees that the vertex cover is within factor γ of the minimum cover, with the quality ratio γ defined as:

$$\gamma = \frac{c}{n - x} \qquad \text{then} \qquad c \leq \gamma c^*$$

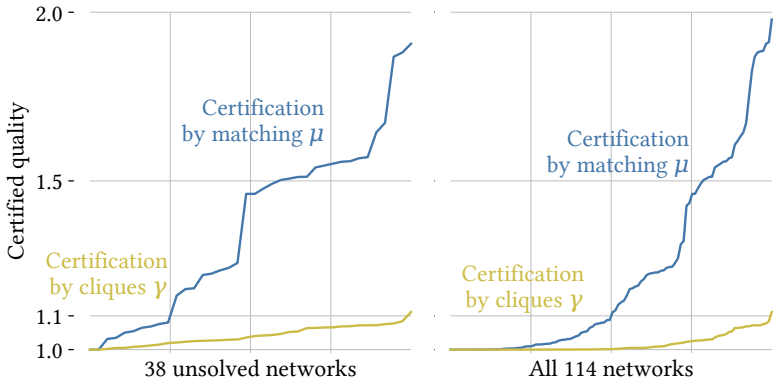

Fig. 1 Ratio of certified quality for vertex cover using matching (μ) or cliques (γ). Left: 38 unsolved networks (exact solution unknown). Right: all 114 networks. Networks are ranked by their certified quality. Horizontal grids indicate 1.1 and 1.5 thresholds and vertical grids cut networks in four even groups. Observe that half of the networks obtain a certified quality within the 1.5 line with matching. With cliques, 92 networks including 18 unsolved have a ratio below 1.03 with cliques: the lower-bound certifies that the smallest cover found by the heuristic is at most 3% larger than minimum.

3 Certifying results on real-world networks

To assess the quality certification method using these bounds, we gather 114 real-world networks (web graphs, social networks, biological interactions...) and we apply the following algorithms on each of them:

- the exponential algorithm of [3] that is able to compute the size c^* of a minimum vertex cover for 76 of the 114 networks in less than six hours;
- the greedy heuristic implemented in [4] to find a small cover of size c;
- our implementation of a linear-time greedy matching algorithm that obtains e close to the e^* of the blossom algorithm (which is not linear and has an average relative improvement under 0.3%);
- our implementation of a linear-time greedy algorithm to partition n nodes into x cliques.

The certification ratio μ obtained with a matching is shown in blue on Figure 1. Unsolved networks are those where the exact algorithm could not compute a minimum vertex cover in six hours; the matching guarantees that, for half of them, the approximate cover is at most 1.5 times as big as the minimum. Among solved networks, the quality ratio is even lower, and the execution of these greedy algorithms can be much faster than the exact exponential algorithm.

As a matching is a special case of clique partition, we know that the clique method gives better results. Indeed, the yellow lines of the figure show that the ratio γ is always under 1.11 even for unsolved networks. Strikingly, it is under 1.01 for 76/114 networks: the minimum value is not known but the certification guarantees that the found cover is at most 1% larger than optimum.

Conclusion

Altogether, the results show that the certification can be used as an efficient shortcut for hard problems: in linear time, we obtain a proof that an approximate result is close to the unknown optimum. Further research needs to translate this principle into applications and to extend it to other algorithmic problem. The hope is that heuristics will always go along with a quality certification method, thus bridging the gap between predictable execution time and guarantees on the results.

References

1. R.M. Karp, *Reducibility among Combinatorial Problems*, Complexity of Computer Computations, 1972. https://doi.org/10.1007/978-1-4684-2001-2_9
2. A. Baudin et. al, *Clique percolation method: memory efficient almost exact communities*, ADMA, 2021. http://arxiv.org/abs/2110.01213
3. D. Hespe et. al, *WeGotYouCovered: the winning solver from the PACE 2019 Challenge*, Workshop CSC, 2020. https://doi.org/10.1137/1.9781611976229.1
4. S. Cai et. al, *Finding a small vertex cover in massive sparse graphs*, Journal of Artificial Intelligence Research, 2017. https://doi.org/10.1613/jair.5443
5. J. Edmonds, *Paths, trees, and flowers*, Canadian Journal of Mathematics, 1965.