Computers and Operations Research 87 (2017) 63–82

Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor

# A Relax-and-Cut framework for large-scale maximum weight connected subgraph problems

Eduardo Álvarez-Miranda[a,c,*], Markus Sinnl[b]

[a] Department of Industrial Engineering, Universidad de Talca, Curicó, Chile
[b] Department for Statistics and Operations Research, University of Vienna, Vienna, Austria
[c] Department of Industrial Engineering, Universidad de Talca, Merced 437, Curicó, Chile

ARTICLE INFO

*Article history:*
Received 1 May 2016
Revised 25 May 2017
Accepted 25 May 2017
Available online 26 May 2017

*Keywords:*
Maximum weight connected subgraph problem
Lagrangian pelaxation
Bioinformatics

ABSTRACT

Finding maximum weight connected subgraphs within networks is a fundamental combinatorial optimization problem both from theoretical and practical standpoints. One of the most prominent applications of this problem appears in Systems Biology and it corresponds to the detection of *active subnetworks* within gene interaction networks.

Due to its importance, several modeling and algorithmic strategies have been proposed for tackling the maximum weight connected subgraph problem (MWCS) over the last years; the most effective strategies typically depend on the use of integer linear programming (ILP). Nonetheless, this implies that large-scale networks (such as those appearing in Systems Biology) can become burdensome; moreover, not all practitioners may have access to an ILP solver. In this paper, a unified modeling and algorithmic scheme is designed to solve the MWCS and some of its application-oriented variants with cardinality-constraints or budget-constraints. The proposed framework is based on a general node-based model which is tackled by a Relax-and-Cut scheme, i.e., Lagrangian relaxation combined with constraint generation; this yields a heuristic procedure capable of providing both dual and primal bounds. The approach is enhanced by additional valid inequalities, lifted valid inequalities, primal heuristics and variable-fixing procedures.

Computational results on instances from the literature, as well as on additional large-scale instances, show that the proposed framework is competitive with respect to the existing approaches and it allows to find improved solutions for some unsolved instances from literature. The effect of initializing a Branch-and-Cut approach with information from the Relax-and-Cut is also investigated. The implemented approach is made available online.



## 1. Introduction and motivation

The problem of finding *active subnetworks* has recently received considerable attention from the bionformatics community (see, e.g., Andreotti, 2015; Backes et al., 2011; Dittrich et al., 2008; El-Kebir, 2015; Hatem, 2014; Huang, 2011; Ideker et al., 2002; Yamamoto et al., 2009 and the references therein). In this problem, one is given a gene interaction network (also known as *interactome*), and the goal is to find active subnetworks associated with a particular biological process (e.g., a cancer). In Dittrich et al. (2008), this problem was formalized as the *maximum weight connected subgraph problem* (MWCS). In this problem, the input corresponds to a graph $G = (V, E)$, with node-weights $w_v \in \mathbb{R}$, $\forall v \in V$, and the goal is to find a connected subgraph $G'$ with maximum node-weight. In a biological context, the nodes model genes and a particular node-weight $w_v$ is a *score* that represents the significance of gene $v$ for the biological process under investigation. The scores are typically based on data obtained by DNA-microarray experiments.

Aside from its importance in bioinformatics, the MWCS appears as a basic optimization problem in wildlife corridor design (Dilkina and Gomes, 2010), forestry planning (Carvajal et al., 2013), object and activity saliency detection (Adluru et al., 2014; Chen and Grauman, 2012; Vijayanarasimhan and Grauman, 2011), wireless network deployment planning (Kuo et al., 2015), among others.

Dittrich et al. (2008) showed that the MWCS can be transformed into the *prize-collecting Steiner tree problem* (PCSTP) and developed an exact integer linear programming (ILP)-based solution approach built on the PCSTP framework of Ljubić et al. (2006). After Dittrich et al. (2008), further exact solution approaches based on ILPs have been proposed by Althaus and Blumenstock (2014),

* Corresponding author.
*E-mail address:* ealvarez@utalca.cl (E. Álvarez-Miranda).

http://dx.doi.org/10.1016/j.cor.2017.05.015
0305-0548/

Álvarez-Miranda et al. (2013a), Álvarez-Miranda et al. (2013b), Backes et al. (2011), El-Kebir and Klau (2014) and Fischetti et al. (2016). In contrast to Dittrich et al. (2008), where arc and node-variables are used, these latter approaches are based on formulations using only node-variables. Such models have much less variables which is particularly useful in highly dense networks as those appearing in the identification of functional modules in gene regulatory networks. Complementary to these algorithms, polyhedral studies on the connected subgraph polytope are carried out in Wang et al. (2017).

Depending on the area of application, there may be additional side-constraints like a cardinality-constraint or a budget-constraint. In Backes et al. (2011), the cardinality-constrained counterpart of the MWCS was tackled via ILP considering an arc-based model. In a latter work, the same variant was approached in Álvarez-Miranda et al. (2013a) using a much more efficient node-based model. Similarly, an arc-based ILP model was proposed by Dilkina and Gomes (2010) for the budget-constrained variant. In a more recent work, considerably better computational results were obtained by means of a node-based model by Álvarez-Miranda et al. (2013b).

In this work, a Relax-and-Cut (R&C) approach, i.e., Lagrangian relaxation combined with constraint generation, is designed for the MWCS, and its cardinality-constrained and budget-constrained versions. The use of such algorithm is motivated by the following two observations. First, the previously proposed ILP approaches need an exponential number of constraints, hence, they are tackled by means of Branch-and-Cut (B&C). As a consequence, these strategies typically fail in providing acceptable gaps for massive instances, as those appearing in bioinformatics, mainly due to time consuming separation procedures. And second, practitioners may not have access to an ILP solver or may not have the expertise to use it, and thus there is need for alternative approaches. As a matter of fact, for this reason the R-package `BioNet` (see Beisser et al., 2010 in addition to Dittrich et al., 2008), also contains a heuristic for the MWCS, for users without access to an ILP solver. Another example of a heuristic for an equivalent problem arising in Bioinformatics corresponds to the hybrid ILP-based heuristic proposed in Akhmedov et al. (2016); the approach is based on embedding the resolution of small size PCSTP instances within a clustering strategy in a divide-and-conquer scheme. In contrast to the heuristics in Akhmedov et al. (2016) and Beisser et al. (2010), the approach proposed in this paper also provides a dual bound which allows to judge the quality of the attained (primal) solutions. Furthermore, the proposed scheme can be embedded within a branch-and-bound framework, allowing an exact resolution of the problem. Note that implementations of Lagrangian relaxation-based algorithms have been successfully applied to solve problems related to the MWCS. Sophisticated Lagrangian relaxation schemes (without cut generation) are designed by Haouari et al. (2008); 2010) for the PCSTP. Complementary, R&C implementations are devised by Lucena (2005); 2006) for the Steiner tree problem, and by Cunha et al. (2009) for the PCSTP. A dual ascent algorithm for the PCSTP is designed in the current working paper (Leitner et al., 2016). This approach also does not need an ILP solver, however, it does not allow for cardinality-constraints or budget-constraints.

In order to assess the efficiency of the R&C algorithm proposed in this paper, both from the view of solution quality and runtime, a computational study on a large set of benchmark instances from the literature is reported. Moreover, additional large-scale instances, which have been constructed to resemble interactomes, are tested as well. For the MWCS, the performance of the R&C is compared with that of the state-of-the-art B&C algorithm proposed in Fischetti et al. (2016). For the cardinality-constrained and budget-constrained MWCS, the proposed algorithm is compared with an adaptation of the same B&C algorithm provided in Fischetti et al. (2016). Furthermore, the use of the R&C as initialization strategy of the B&C algorithm is also investigated. Computational results show the advantages of the R&C algorithm with respect to the other exact alternatives. The implemented program provided for download at https://msinnl.github.io/ (Sinnl and Álvarez-Miranda, 2017).

*Paper outline.* In Section 2 the ILP formulation used in the R&C algorithm is presented. Likewise, the cardinality-constrained and budget-constrained versions are discussed in more detail. A generic scheme of R&C is outlined in Section 3. In Section 4 the designed algorithmic framework is described. Computational results are reported in Section 5, which also gives a description of the B&C, and of the combination of R&C and B&C. Finally, concluding remarks are drawn in Section 6.

## 2. An ILP formulation for the MWCS and some of its variants

A formal definition of MWCS is:

**Definition 1** (The MWCS)**.** Given an undirected graph $G = (V, E)$ and a weight function $\mathbf{w} : V \to \mathbb{R}$, the MWCS is the problem of finding a subset of nodes $V_T$, so that the subgraph $(V_T, E_T)$, with $E_T = \{\{u, v\} \in E \mid u, v \in V_T\}$, is connected and has the largest possible sum of node weights (i.e., $\sum_{v \in V_T} w_v$ is maximized).

As mentioned before, this definition can be complemented by so-called *side-constraints* depending on the particular application. For instance, in some contexts, a constraint $\sum_{v \in V_T} c_v \leq B$ defined by $\mathbf{c} : V \to \mathbb{N}$ and $B \in \mathbb{N}$ must also be satisfied. Such a constraint may appear, e.g., in Bioinformatic settings where *compact*, i.e., cardinality-constrained, functional modules are preferred over large ones (see, e.g.,Yamamoto et al., 2009; Yosef et al., 2011); in this case $\mathbf{c} = \mathbf{1}$. Likewise, in a wildlife corridor design setting although the aim is to find a connected reserve that maximizes the ecological suitability, it must respect an economical bound (Dilkina and Gomes, 2010). Similarly, in the design of wireless networks, although the objective is to construct a mesh that maximizes the service coverage, there are construction budgets that must be satisfied (Kuo et al., 2015). Although the cardinality-constrained version is a special case of the budget-constrained version, here it is regarded as its own problem variant, as it allows to use a more efficient solution approach. In the budget-constrained version, one assumes that $c_i \leq B$, $\forall i \in V$, as nodes not fulfilling this condition can easily be removed at the beginning.

Let $\mathbf{y} \in \{0, 1\}^{|V|}$ be a vector of binary variables such that $y_i = 1$ if node $i \in V$ is part of the connected subgraph, and $y_i = 0$ otherwise. Let $\Phi$ denote the set of all $\{0, 1\}^{|V|}$ vectors associated with connected components of $G$. Under this notation, the MWCS and its variants above can be modeled as

$$\max \left\{ \mathbf{w}^T \mathbf{y} \mid \Lambda \mathbf{y} \leq \boldsymbol{\beta} \text{ and } \mathbf{y} \in \Phi \right\}, \tag{1}$$

where $\Lambda \mathbf{y} \leq \boldsymbol{\beta}$ represents a (possibly empty) set of side-constraints. There are several alternatives to model the constraint $\mathbf{y} \in \Phi$. In this paper a node-based model, as proposed in Álvarez-Miranda et al. (2013a), El-Kebir and Klau (2014) and Fischetti et al. (2016), is considered. For formulating such model, the following definition is needed.

**Definition 2** (Node-separator)**.** For two distinct nodes $k$ and $\ell$ from $V$, a subset of nodes $N \subseteq V \setminus \{k, \ell\}$ is called $(k, \ell)$-*node separator* if and only if after eliminating $N$ from $V$ there is no $(k, \ell)$ path in $G$. A separator $N$ is *minimal* if $N \setminus \{i\}$ is not a $(k, \ell)$ separator, for any $i \in N$. Let $\mathcal{N}(k, \ell)$ denote the family of all minimal $(k, \ell)$ separators.