




Hardness, approximability, and fixed-parameter tractability of the clustered shortest-path tree problem

Mattia D’Emidio¹  · Luca Forlizzi¹ · Daniele Frigioni¹ · Stefano Leucci² · Guido Proietti^{1,3}

Published online: 2 January 2019

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Given an n -vertex non-negatively real-weighted graph G , whose vertices are partitioned into a set of k clusters, a *clustered network design problem* on G consists of solving a given network design optimization problem on G , subject to some additional constraints on its clusters. In particular, we focus on the classic problem of designing a *single-source shortest-path tree*, and we analyse its computational hardness when in a feasible solution each cluster is required to form a subtree. We first study the *unweighted* case, and prove that the problem is NP-hard. However, on the positive side, we show the existence of an approximation algorithm whose quality essentially depends on few parameters, but which remarkably is an $O(1)$ -approximation when the largest out of all the *diameters* of the clusters is either $O(1)$ or $\Theta(n)$. Furthermore, we also show that the problem is *fixed-parameter tractable* with respect to k or to the number of vertices that belong to clusters of size at least 2. Then, we focus on the *weighted* case, and show that the problem can be approximated within a tight factor of $O(n)$, and that it is fixed-parameter tractable as well. Finally, we analyse the unweighted *single-pair shortest path problem*, and we show it is hard to approximate within a (tight) factor of $n^{1-\epsilon}$, for any $\epsilon > 0$.

Keywords Clustered shortest-path tree problem · Hardness · Approximation algorithms · Fixed-parameter tractability · Network design

The results presented in this work have been announced in a preliminary form in D’Emidio et al. (2016). This research has partially supported by the Italian National Group for Scientific Computation GNCS-INdAM.

✉ Mattia D’Emidio
mattia.demidio@univaq.it

Extended author information available on the last page of the article

1 Introduction

In several modern network applications, the underlying set of nodes may be partitioned into *clusters*, with the intent of modeling some aggregation phenomena taking place among similar entities in the network. In particular, this happens in communication and social networks, where clusters may refer to local-area subnetworks and to communities of individuals, respectively. While on the one hand the provision of clusters allows to represent the complexity of reality, on the other hand it may ask for introducing some additional constraints on a feasible solution to a given network design problem, with the goal of preserving a specific cluster-based property. Thus, on a theoretical side, given a (possibly weighted) graph G , whose vertex set is partitioned into k pairwise disjoint subsets (i.e., clusters), a *clustered* (a.k.a. *generalized*) *network design problem* on G consists of finding a (possibly optimal) solution to a given network design problem on G , subject to some additional constraints on its clusters. Depending on such constraint, the computational complexity of the resulting problem may change drastically as compared to the unconstrained version. Therefore, this class of problems deserves a theoretical investigation that, quite surprisingly, seems to be rather missing up to now.

One of the most intuitive constraints one could imagine is that of maintaining some sort of *proximity* relationship among nodes in a same cluster. This scenario has immediate practical motivations: for instance, in a communication network, this can be convincingly justified with the requirement of designing a network on a classic two-layer (i.e., local *versus* global layer) topology. In particular, if the foreseen solution has to be a (spanning) tree T in G , then a natural requirement is that each cluster should induce a (connected) subtree of T . For the sake of simplicity, in the following this will be referred to as a *clustered tree design problem* (CTDP), even if this is a slight abuse of terminology. Correspondingly, classic spanning-tree optimization problems on graphs can be reconsidered under this new perspective, aiming at verifying whether they exhibit a significant deviation (from a computational point of view) w.r.t. the ordinary (i.e., non-clustered) counterpart. In particular, we will focus on the clustered version of the problem of computing a *single-source shortest-path tree* (SPT) of G , i.e., a spanning tree of G rooted at a given source node, say s , minimizing the total length of all the paths emanating from s . It is worth noticing that an SPT, besides its theoretical relevance, has countless applications, and in particular it supports a set of primitives of primary importance in communication networks, as for instance the *broadcast protocol* and the *spanning tree protocol*.

1.1 Contribution of the paper

Let $G = (V, E, w)$ be a connected and undirected graph of n vertices and m edges, where each edge $(u, v) \in E$ is associated with a non-negative real weight $w(u, v)$. For a subgraph H of G , we will use $V(H)$ ($E(H)$, resp.) to denote the set of vertices (edges, resp.) of H , and $H[S]$ to denote the subgraph of H induced by S , $S \subseteq V(H)$. Moreover, $\pi_H(u, v)$ will denote a *shortest path* between vertices u and v in H , while

$d_H(u, v)$ will denote the corresponding *distance* between u and v in H , i.e., the sum of the weights of the edges in $\pi_H(u, v)$.

Formally, the clustered version of the SPT problem (CLUSPT in the sequel), is defined as follows. We are given a graph G defined as above, along with a partition of V into a set of k (pairwise disjoint) clusters $\mathcal{V} = \{V_1, V_2, \dots, V_k\}$, and a distinguished source vertex $s \in V$. The objective is to find a *clustered SPT* of G rooted at s , i.e., a spanning tree T of G such that $T[V_i]$, $i = 1, \dots, k$, is a connected component (i.e., a subtree) of T , and for which the *broadcast cost* from s , i.e. $\text{COST}(T) = \sum_{v \in V} d_T(s, v)$, is minimum.

The SPT problem in a non-clustered setting has been widely studied, and in its more general definition it can be solved in $O(m + n \log n)$ time by means of the classic Dijkstra's algorithm. More efficient solutions are known for special classes of graphs (e.g., euclidean, planar, directed acyclic graphs, etc.), or for restricted edge weights instances. In particular, if w is uniform, namely G is *unweighted*, then an optimal solution can be found in $O(m + n)$ time by means of a simple *breadth-first search* (BFS) visit of G . Nevertheless, to the best of our knowledge nothing is known about its clustered variant, despite the fact that, as we argued above, it is very reasonable to imagine a scenario where an efficient broadcast needs to be applied locally and hierarchically within each cluster.

Here, we then try to fill this gap, and we show that CLUSPT, and its unweighted version, say CLUBFS, are actually much harder than their standard counterparts, namely:

1. CLUBFS is NP-hard, but it admits an $O(\min\{\frac{4nk}{\gamma}, \frac{4n^2}{\gamma^2}, 2\gamma\})$ approximation algorithm, where γ denotes the length of the largest out of all the *diameters* of the clusters. Interestingly, the approximation ratio becomes $O(1)$ when γ is either $O(1)$ or $\Theta(n)$, which may cover cases of practical interest. However, we also point out that in the worst case, namely for $\gamma = \Theta(n^{\frac{2}{3}})$ and $k = \Theta(\sqrt[3]{n})$, the algorithm becomes $O(n^{\frac{2}{3}})$ -approximating. Besides that, we also show that the problem is *fixed-parameter tractable*, as we can provide a $\tilde{O}\left(\min\left\{2^k k^3 n^4, h^{\frac{h}{2}} m\right\}\right)$ time exact algorithm,¹ where h is the total number of vertices that belong to clusters of size at least two.
2. CLUSPT is hard to approximate within a factor of $n^{1-\epsilon}$ for any constant $\epsilon \in (0, 1]$, unless $P = NP$, but, on the positive side: (i) it admits an n -approximation, thus essentially tight, algorithm; (ii) similarly to the unweighted case, it is fixed-parameter tractable as well.

Finally, we study the *clustered single-pair shortest path problem* (say CLUSP in the sequel) on unweighted graphs, i.e., the problem of finding a shortest path between a given pair of vertices of G , subject to the constraint that the vertices from a same cluster that belong to the path must appear consecutively. Notice that in this variant, not *all* the vertices of a cluster must belong to a solution, and not *all* the clusters must enter into a solution. We show that it cannot be approximated in polynomial time within a factor of $n^{1-\epsilon}$, for any constant $\epsilon > 0$, unless $P = NP$. This extends the inapproximability result (within any polynomial factor) that was given in Lin and Wu

¹ Throughout the paper, the notation \tilde{O} suppresses factors that are polylogarithmic in n .

(2016) for the corresponding weighted version. Since obtaining an n -approximation is trivial, the provided inapproximability result is a bit surprising, as one could have expected the existence of a $o(n)$ -approximation algorithm, similarly to what happened for CLUBFS.

1.2 Related work

Several classic tree/path-design problems have been investigated in the CTDP framework. Some of them, due to their nature, do not actually exhibit a significant deviation (from a computational point of view) w.r.t. the ordinary (i.e., non-clustered) counterpart. For instance, the *minimum spanning tree* (MST) problem falls in this category, since we can easily solve its clustered version by first computing a MST of each cluster, then contracting these MSTs each to a vertex, and finally finding a MST of the resulting graph. This favourable behaviour is an exception, however, as the next cases show.

A well-known clustered variant of the *traveling salesperson problem* is that in which one has to find a minimum-cost Hamiltonian cycle of G (where G is assumed to be complete, and w is assumed to be a metric on G) such that *all* the vertices of each cluster are visited consecutively. For this problem, Bao and Liu (2012) give a $13/6$ -approximation algorithm, thus improving a previous approximation ratio of 2.75 due to Guttmann-Beck et al. (2000). As a comparison, recall that the best old-standing approximation ratio for the unclustered version of the problem is equal to $3/2$ (i.e., the celebrated Christofides algorithm).

Another prominent clustered variant is concerned with the classic *minimum Steiner tree problem*. In this case, one has to find a tree of minimum cost spanning a subset $R \subseteq V$ of *terminal* vertices, under the assumption that nodes in R are partitioned into a set of clusters, say $\{R_1, R_2, \dots, R_k\}$, and with the additional constraint that, in a feasible solution T , we have that, for every $i = 1, 2, \dots, k$, the minimal subtree of T spanning the vertices of R_i does not contain any terminal vertex outside R_i . For this problem, again restricted to the case in which G is complete and w is a metric on G , in Wu and Lin (2015) the authors present a $(2 + \rho)$ -approximation algorithm, where $\rho \simeq 1.39$ is the best known approximation ratio for the minimum Steiner tree problem (Byrka et al. 2013).

We also mention the clustered variant of the *minimum routing-cost spanning tree problem*. While in the non-clustered version one has to find a spanning tree of G minimizing the sum of all-to-all tree distances, and the problem is known to admit a PTAS (Wu et al. 1998), in Lin and Wu (2016) the authors analyze the clustered version, and show that on general graphs the problem is hard to approximate within any polynomial factor, while if G is complete and w is a metric on G , then the problem admits a factor-2 approximation. Interestingly, along the way the authors present an inapproximability result for CLUSP (on weighted graphs), which was in fact the inspiration for the present study.

Further, we refer the reader to the paper by Feremans et al. (2003), where the authors review several classic network design problems in a clustered perspective, but with different side constraints on the clusters (i.e., expressed in terms of number of

representatives for each cluster that has to belong to a feasible solution). A notable example of this type is the *group Steiner tree problem*, where it is required that *at least* one terminal vertex from each cluster R_i must be included in a feasible solution. This problem is known to be approximable within $O(\log^3 n)$ (Garg et al. 2000), and not approximable within $\Omega(\log^{2-\epsilon} n)$, for any $\epsilon > 0$, unless NP admits quasipolynomial-time Las Vegas algorithms (Halperin and Krauthgamer 2003).

Finally, we remark that the problem of deciding whether there exists a set of clusters that satisfies some property of interest has also been extensively studied (e.g., a typical requirement in social networks is for clusters, which represent communities, to induce dense connected subgraphs), as well as the dual problem of editing a graph so that it contain a set of clusters with the sought property. We refer the interested reader to Zou et al. (2018) and to the references therein.

1.3 Structure of the paper

For the sake of clarity, we first present, in Sect. 2, our results on CLUBFS. Then, in Sect. 3, we give our results on CLUSPT, while in Sect. 4 we discuss our results on unweighted CLUSP. Finally, in Sect. 5 we conclude the paper and outline possible future research directions.

2 CLUBFS

In this section, we present our results on the CLUBFS problem. In particular, we first prove that it is NP-hard, then we show that it can be approximated within an $O(n^{\frac{2}{3}})$ -factor in polynomial time, by providing a suitable approximation algorithm, and finally we show it is fixed-parameter tractable. We start by proving the following result:

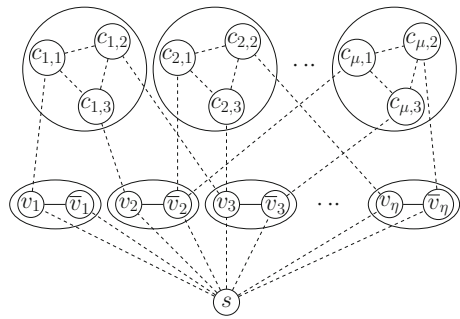
Theorem 1 CLUBFS is NP-hard.

Proof In order to prove the statement we provide a polynomial-time reduction from the 3-CNF-SAT problem, which is known to be NP-complete, to CLUBFS.

The 3-CNF-SAT problem is a variant of the classic CNF-SAT problem. CNF-SAT is the problem of determining whether it is satisfiable a given *Boolean CNF formula*, i.e., a conjunction of *clauses*, where a clause is a disjunction of *literals*, and a literal represents either a *variable* or its negation. In the 3-CNF-SAT version, the number of literals in each clause is constrained to be exactly three.

The proof proceeds as follows: starting from a 3-CNF-SAT instance ϕ with η variables, say x_1, \dots, x_η , and μ clauses, say c_1, \dots, c_μ , we first construct an instance $\langle G_\phi, \mathcal{V}, s \rangle$ of the CLUBFS problem which consists of: (i) a graph G_ϕ ; (ii) a clustering \mathcal{V} of the vertices of G_ϕ ; (iii) a distinguished source vertex s of G_ϕ . We then show that instance $\langle G_\phi, \mathcal{V}, s \rangle$ exhibits the two following properties: (i) if ϕ is satisfiable then $\text{OPT} \leq 3\eta + 8\mu$; (ii) if ϕ is not satisfiable then $\text{OPT} \geq 3\eta + 8\mu + 3$, where OPT denotes the cost of the optimal solution to the CLUBFS problem on $\langle G_\phi, \mathcal{V}, s \rangle$. By proving the above, we will show that finding an optimal solution to the CLUBFS problem is at least as hard as solving 3-CNF-SAT.

Fig. 1 Graphical representation of the reduction from 3-CNF-SAT to CLUBFS used in the proof of Theorem 1



The graph G_ϕ corresponding to the formula ϕ can be obtained from an empty graph by proceeding as follows. First, we add to $V(G_\phi)$ a single source vertex s and for each variable x_i , we add: (i) two *variable vertices* v_i and \bar{v}_i to $V(G_\phi)$; (ii) three edges, namely (v_i, s) , (\bar{v}_i, s) and (v_i, \bar{v}_i) , to $E(G_\phi)$. Then, for each clause c_j we add: (i) three *clause vertices*, $c_{j,1}, c_{j,2}, c_{j,3}$, one for each of the three literals of c_j , to $V(G_\phi)$; (ii) three edges $(c_{j,1}, c_{j,2}), (c_{j,2}, c_{j,3}), (c_{j,3}, c_{j,1})$ to $E(G_\phi)$. Finally, for each clause c_j , and for $k = 1, 2, 3$, let x_i be the variable associated with the k th literal ℓ of c_j . If the literal is negative, i.e., $\ell = \bar{x}_i$, we add edge $(c_{j,k}, \bar{v}_i)$ to $E(G_\phi)$, otherwise (i.e., $\ell = x_i$) we add $(c_{j,k}, v_i)$ to $E(G_\phi)$.

It is easy to see that G_ϕ has $|V(G_\phi)| = 3\mu + 2\eta + 1$ vertices and $|E(G_\phi)| = 6\mu + 3\eta$ edges. A clarifying example on how to build G_ϕ for a 3-CNF-SAT formula ϕ is shown in Fig. 1. Notice, e.g., that the first literal of clause c_1 is positive and associated with variable x_1 . Therefore, clause vertex $c_{1,1}$ is connected to variable vertex v_1 in G_ϕ .

Now, the final step of the construction consists in defining a clustering \mathcal{V} over the vertices of G_ϕ . In details, we define $\mathcal{V} = \{V_s, V_1, \dots, V_\mu, V_{\mu+1}, V_{\mu+\eta}\}$ as follows. The source vertex is a singleton, i.e., V_s contains s only. Then, for each clause c_j , with $j = 1, \dots, \mu$, we set $V_j = \{c_{j,1}, c_{j,2}, c_{j,3}\}$. Finally, for each variable x_i we set $V_{\mu+i} = \{\bar{v}_i, v_i\}$.

We now proceed with the last part of the proof. In particular, if ϕ is satisfiable, we consider a satisfying assignment and we construct a solution T to CLUBFS on instance $\langle G_\phi, \mathcal{V}, s \rangle$ as follows: i) for each variable x_i , if x_i is true we add the edges (s, v_i) and (v_i, \bar{v}_i) to T , otherwise we add the edges (s, \bar{v}_i) and (v_i, \bar{v}_i) to T ; ii) for each clause c_j , choose $k \in \{1, 2, 3\}$ so that the k th literal of c_j is true, and let v_i be the unique variable vertex that is a neighbor of $c_{j,k}$ in G_ϕ . We add the edges in $\{(c_{j,k}, v_i)\} \cup \{(c_{j,k}, c_{j,k'}) : k' \in \{1, 2, 3\} \wedge k' \neq k\}$ to T .

It is easy to check that exactly one of each pair of vertices v_i and \bar{v}_i is at distance 1 from s in T while the other is at distance 2. Moreover, for each clause c_j exactly one of the vertices in $\{c_{j,1}, c_{j,2}, c_{j,3}\}$ is at distance 2 from s in T , while the other two are at distance 3. Hence $\text{OPT} \leq 3\eta + 8\mu$. Suppose now that ϕ is not satisfiable and let T be a solution to CLUBFS. It is easy to see that, for each variable x_i , solution T must include the edge (v_i, \bar{v}_i) since the graph induced by the associated cluster must be connected. This means that at least one of v_i and \bar{v}_i must be at distance 2 from s in T . Similarly, since for every $j = 1, \dots, \mu$ the subgraph of T induced by the vertices in $\{c_{j,1}, c_{j,2}, c_{j,3}\}$ must be connected, we have that one of them, say w.l.o.g.

$c_{j,1}$, must be at a distance at least $d_T(s, c_{j,1}) \geq d_{G_\phi}(s, c_{j,1}) = 2$ from s while the other two must be at a distance of at least $d_T(s, c_{j,1}) + 1$ in T . Moreover, since ϕ is not satisfiable, there is at least one clause c_j with $j \in \{1, \dots, \mu\}$ such that the closest vertex of $c_{j,1}, c_{j,2}, c_{j,3}$ is at distance at least 3 from s in T . Indeed, if that were not the case, this would imply that, for each clause c_j , there would exist a vertex $c_{j,k}$, for a certain k , at distance 2 from s in T , and hence the set of vertices at distance 1 from s would induce a satisfying truth assignment for ϕ . It follows that:

$$\begin{aligned} \text{COST}(T) &\geq \eta + 2\eta + (\mu - 1)2 + 2(\mu - 1)3 + 3 + 2 \cdot 4 \\ &= 3\eta + 8(\mu - 1) + 11 \\ &= 3\eta + 8\mu + 3. \end{aligned}$$

Since the latter bound holds for any solution to CLUBFS, we have $\text{OPT} \geq 3\eta + 8\mu + 3$, which concludes the proof. \square

2.1 An approximation algorithm

In what follows, we present an approximation algorithm for CLUBFS (see Algorithm 1). The main idea of the algorithm is that of minimizing the number of distinct clusters that must be traversed by any path connecting the source s to a vertex $v \in V$. Recall that the *diameter* $\text{DIAM}(G)$ of a graph G is the length of a longest shortest path in G . Then, it is possible to show that: (i) if all the clusters are of low diameter, then this leads to a good approximation for CLUBFS; (ii) otherwise, i.e. if at least one cluster has large diameter, then the optimal solution must be expensive and hence any solutions for CLUBFS will provide the sought approximation.

Given an instance $\langle G, \mathcal{V}, s \rangle$ of CLUBFS, w.l.o.g. let us assume that V_1 is the cluster containing vertex s , and that $G[V_i]$ is connected for each $i = 1, \dots, k$, as otherwise the problem trivially admits no feasible solution. Our approximation algorithm first considers each cluster $V_i \in \mathcal{V}$ and identifies all the vertices belonging to V_i into a single *cluster-vertex* v_i to obtain a graph G' in which: (i) each vertex corresponds to a cluster; (ii) there is an edge (v_i, v_j) between two vertices in G' if and only if the set $E_{i,j} = \{(v_i, v_j) \in E : v_i \in V_i \wedge v_j \in V_j \wedge i \neq j\}$ is not empty. The algorithm proceeds then by computing a BFS tree T' of G' rooted at v_1 and constructs the sought approximate solution \tilde{T} as follows: initially, \tilde{T} contains all the vertices of G and the edges of a BFS tree of $G[V_1]$ rooted at s . Then, for each edge (v_i, v_j) of T' , where v_i is the parent of v_j in T' , it adds to \tilde{T} a single edge $(v_i, r_j) \in E_{i,j}$ along with all the edges of a BFS tree T_j of $G[V_j]$ rooted at r_j .

We now show that Algorithm 1 outputs a feasible solution for the CLUBFS problem which is far from the optimum by at most a factor of $\min\{\frac{4nk}{\gamma}, \frac{4n^2}{\gamma^2}, 2\gamma\}$, where $\gamma = \max_{V_i \in \mathcal{V}} \text{DIAM}(G[V_i])$. In particular, given an instance $\langle G, \mathcal{V}, s \rangle$ of CLUBFS, let T^* be an optimal clustered BFS tree. To prove the approximation ratio, we will make use of the following lemma.

Lemma 1 $\text{COST}(\tilde{T}) \leq 2\gamma \text{COST}(T^*)$.

Algorithm 1: Approximation algorithm for CLUBFS.

```

Input : An instance  $(G, \mathcal{V}, s)$  of CLUBFS
Output: An approximated clustered BFS tree  $\tilde{T}$ 
1 Let  $s \in V_1$  w.l.o.g.
2  $G' \leftarrow$  Copy  $G$  and identify the vertices belonging to  $V_i$  into a single cluster-vertex  $v_i$ 
3 Let  $E_{i,j} = \{(v_i, v_j) \in E : v_i \in V_i \wedge v_j \in V_j\}$ 
4  $T' \leftarrow$  Compute a BFS tree of  $G'$  rooted at  $v_1$ 
5  $T_1 \leftarrow$  Compute a BFS tree of  $G[V_1]$  rooted at  $s$ 
6  $\tilde{T} \leftarrow (V, E(T_1))$ 
7 for  $j = 2, \dots, k$  do
8    $(p_i, r_j) \leftarrow$  any edge in  $E_{i,j}$  where  $v_i$  is the parent of  $v_j$  in  $T'$ 
9    $T_j \leftarrow$  Compute a BFS tree of  $G[V_j]$  rooted at  $r_j$ 
10   $E(\tilde{T}) \leftarrow E(\tilde{T}) \cup \{(p_i, r_j)\} \cup E(T_j)$ 
11 return  $\tilde{T}$ 
    
```

Proof We first prove that for every $v \in V$ it holds that $d_{\tilde{T}}(s, v) \leq \gamma(d_{T^*}(s, v) + 1)$. In particular, let us assume that V_i is the cluster of \mathcal{V} containing v . Moreover, let T'' be the tree obtained from T^* by identifying each cluster $V_j \in \mathcal{V}$ into a single cluster-vertex τ_j . Observe that r_i is the vertex chosen by Algorithm 1 at line 8 w.r.t. the cluster V_i containing v . Therefore, we have that:

$$\begin{aligned}
 d_{\tilde{T}}(s, v) &= d_{\tilde{T}}(s, r_i) + d_{\tilde{T}}(r_i, v) \leq \gamma d_{T'}(s, v_i) + \gamma \\
 &\leq \gamma d_{T''}(s, \tau_i) + \gamma \leq \gamma d_{T^*}(s, v) + \gamma \\
 &= \gamma (d_{T^*}(s, v) + 1),
 \end{aligned}$$

from which it follows:

$$\begin{aligned}
 \text{COST}(\tilde{T}) &= \sum_{v \in V} d_{\tilde{T}}(s, v) \leq \sum_{v \in V} \gamma (d_{T^*}(s, v) + 1) \\
 &\leq \gamma \sum_{v \in V} d_{T^*}(s, v) + \gamma n
 \end{aligned}$$

and therefore $\text{COST}(\tilde{T}) \leq \gamma \text{COST}(T^*) + \gamma n \leq 2\gamma \text{COST}(T^*)$. □

Given the above lemma, we are now ready to prove the following theorem.

Theorem 2 Algorithm 1 is a polynomial-time ρ -approximation algorithm for CLUBFS, where $\rho = \min\{\frac{4nk}{\gamma}, \frac{4n^2}{\gamma^2}, 2\gamma\}$.

Proof First of all, note that there is a least one cluster V_i such that $\text{DIAM}(G[V_i]) = \gamma$, and hence it follows that $\text{COST}(T^*) \geq \frac{\gamma^2}{4}$. Indeed, if γ is even, we have that an optimal solution must pay at least the cost of two paths rooted at the center of a diametral path, namely

$$\text{COST}(T^*) \geq 2 \sum_{i=1}^{\gamma/2} i = \frac{\gamma^2}{4} + \frac{\gamma}{2}.$$

Similarly, if γ is odd, we have

$$\text{COST}(T^*) \geq \sum_{i=1}^{(\gamma-1)/2} i + \sum_{i=1}^{(\gamma+1)/2} i = \frac{\gamma^2}{4} + \frac{\gamma}{2} + \frac{1}{4}.$$

Now, we observe that $\text{COST}(\tilde{T})$ is upper bounded by:

- (i) γnk , since in any feasible solution T to CLUBFS, it holds that $d_T(s, v) \leq \gamma k, \forall v \in V$;
- (ii) n^2 , since $d_G(s, v) \leq n, \forall v \in V$.

Therefore, since $\text{COST}(T^*) \geq \frac{\gamma^2}{4}$, the approximation ratio achieved by Algorithm 1 is always upper bounded by $\min\{\gamma nk, n^2\} \cdot \frac{4}{\gamma^2} = \min\{\frac{4nk}{\gamma}, \frac{4n^2}{\gamma^2}\}$. Moreover, by Lemma 1 we also know that $\text{COST}(\tilde{T}) \leq 2\gamma \text{COST}(T^*)$. Hence, overall, Algorithm 1 always computes a solution \tilde{T} such that $\frac{\text{COST}(\tilde{T})}{\text{COST}(T^*)} \leq \rho$, where $\rho = \min\{\frac{4nk}{\gamma}, \frac{4n^2}{\gamma^2}, 2\gamma\}$. Since the time complexity is upper bounded by the cost of computing the BFS trees, the claim follows. □

Notice that each of the three terms in ρ can be the minimum one, depending on the structure of a given instance of CLUBFS. In particular, the first term is the unique minimum when $\sqrt{2nk} < \gamma < \frac{n}{k}$, and the considered interval is not empty, i.e., when $\sqrt{2nk} < \frac{n}{k}$, which implies $k < \sqrt[3]{n/2}$. In this latter case, the second term is to be preferred when $\gamma > \frac{n}{k}$, while the minimum is attained by the third term when $\gamma < \sqrt{2nk}$. Otherwise, i.e., when $k \geq \sqrt[3]{n/2}$, and hence the aforementioned interval is empty, then the second (resp., third) term is the unique minimum when γ is larger (resp., smaller) than $\sqrt[3]{2n^2}$. Remarkably, when γ is either $O(1)$ or $\Theta(n)$, our algorithm thus provides an $O(1)$ -approximation ratio. Finally, notice that if we set $\gamma = \Theta(n^{\frac{2}{3}})$ and $k = \Theta(n^{\frac{1}{3}})$, then the three terms in ρ coincide and are equal to $\Theta(n^{\frac{2}{3}})$, which then happens to be the achieved ratio of our approximation algorithm in the worst case.

2.2 Fixed-parameter tractability

In this section, we prove that CLUBFS is *fixed-parameter tractable* (FPT) w.r.t. two natural cluster-related parameters, by providing two different FPT algorithms, namely CLUBFS–FPT1 and CLUBFS–FPT2. Recall that an FPT algorithm is allowed to have an exponential running time, but only in terms of some natural parameter of the problem instance that can be expected to be small in typical applications.

2.2.1 Algorithm CLUBFS–FPT1

In CLUBFS–FPT1 we choose as our first natural parameter the number k of clusters of \mathcal{V} . Notice that every feasible solution T for CLUBFS induces a *cluster-tree* T_C obtained from T by identifying the vertices of the same cluster into a single vertex. The main

idea underlying the algorithm is that of guessing, for each vertex of an optimal cluster-tree T_C^* , the vertices belonging to the subtrees rooted in (one of) its children and then to iteratively reconstruct T_C^* . For the sake of simplicity, in the following we will assume that n is a power of two. However, note that this assumption can be removed by either modifying the input graph or by tweaking the definition of the functions $f_{v,i}$ and $g_{v,i}$ that are given later in this section.

We start with some definitions. First, given an instance $\langle G, \mathcal{V}, v \rangle$ of CLUBFS, for any $V_i \in \mathcal{V}$, and $v \in V_i$, we let $\text{BFS}_{V_i}[v]$ be the cost of a BFS tree of $G[V_i]$ having source vertex v , i.e.,

$$\text{BFS}_{V_i}[v] = \sum_{u \in V_i} d_{G[V_i]}(v, u).$$

Then, we define a set $U = \mathcal{V} \cup A$ as the union of the set of clusters \mathcal{V} with a set $A = \{a_1, \dots, a_{\log n}\}$, containing $\log n$ additional elements. Moreover, we let $\mu : 2^A \rightarrow V$ be a bijection that maps each of the $2^{\log n} = n$ subsets of A to a vertex of V .

For each $H \subseteq U$, we let $\text{OPT}_{v,i}[H]$ be a quantity depending on the cost c^* of an optimal solution to an auxiliary instance $\langle G', H \cap \mathcal{V}, v \rangle$ of CLUBFS, where G' is the subgraph of G induced by the vertices in $\cup_{C \in H \cap \mathcal{V}} C$, provided that the following constraints are all satisfied:

- (i) $|H \cap \mathcal{V}| \leq i$ (i.e., we restrict to subproblems having at most i clusters);
- (ii) $v \in \cup_{C \in H \cap \mathcal{V}} C$;
- (iii) G' is connected;
- (iv) $A \subseteq H$.

Let M be a parameter whose value will be specified later. If all the above mentioned four constraints are satisfied and $c^* < M$, then we define $\text{OPT}_{v,i}[H] = c^*$. Otherwise, if (i) is satisfied and either $c^* \geq M$ or at least one of (ii)–(iv) is not satisfied, then we allow $\text{OPT}_{v,i}[H]$ to be any value larger than or equal to M . Finally, if (i) is not satisfied, then we allow $\text{OPT}_{v,i}[H]$ to be any upper bound to c^* .

Therefore, according to our definition, we have that, whenever (i), (ii), (iii), and (iv) are satisfied, we can set:

$$\text{OPT}_{v,1}[H] = \min\{\text{BFS}_{H \cap \mathcal{V}}[v], M\} \tag{1}$$

Otherwise we set $\text{OPT}_{v,1}[H] = M$.

Now, let $\eta(H \cap \mathcal{V}) = \sum_{C \in H \cap \mathcal{V}} |C|$ be the number of vertices in the clusters of $H \cap \mathcal{V}$. Moreover, given a vertex $v \in V$, let

$$\ell(v, v') = \min_{\substack{(x,v') \in E(G) \\ x \in V(G[R])}} \{d_{G[R]}(v, x) + 1\},$$

where $R \in H \cap \mathcal{V}$ is the cluster containing v and $\ell(v, v')$ is the shortest among the paths from v to v' that traverse only vertices in R , except for v' . If there is no such path, then $\ell(v, v') = +\infty$.

Hence, for $i > 1$ we can write the following recursive formula:

$$\begin{aligned} \text{OPT}_{v,i}[H] = \min_{H' \subseteq H} \{ & L(v, \mu(H' \cap A), H' \cap V) \\ & + \text{OPT}_{\mu(H' \cap A), i-1}[(H' \cap \mathcal{V}) \cup A] \\ & + \text{OPT}_{v, i-1}[(H \setminus H') \cup A], \end{aligned} \tag{2}$$

where $L(v, \mu(H' \cap A), H' \cap V) = \min\{\ell(v, v')\eta(H' \cap \mathcal{V}), M\}$ accounts for (the lengths of) the portions of the shortest paths from v to the vertices in $C' = \cup_{C \in H' \cap V} C$ whose edges are not in the subgraph induced by C' .

Given the above formula, we now show that $\text{OPT}_{v,i}[H]$ for $i > 1$ can be computed efficiently by exploiting a result provided in Björklund et al. (2007), namely the following:

Theorem 3 (Björklund et al. 2007) *Given a set X and two functions $f, g : 2^X \rightarrow [-W, \dots, W]$, it is possible to compute in $\text{conv}(W, X) := O(W \cdot |X|^3 \cdot 2^{|X|} \cdot \text{polylog}(W, |X|))$ time² the subset convolution $(f * g)$ of f and g over the min-sum semiring, i.e., for every set $Y \subseteq X$ the quantity:*

$$(f * g)(Y) = \min_{Z \subseteq Y} \{f(Z) + g(Y \setminus Z)\}.$$

In particular, the main idea here is to express the values $\text{OPT}_{v,i}[H]$ as subset convolutions of two suitable functions f and g . In more details, notice that Eq. (2) can be rewritten as follows:

$$\text{OPT}_{v,i}[H] = \min_{H' \subseteq H} \{f_{v,i}(H') + g_{v,i}(H \setminus H')\} = (f_{v,i} * g_{v,i})(H) \tag{3}$$

once we define

$$\begin{aligned} f_{v,i}(X) &= L(v, \mu(X \cap A), X \cap V) + \text{OPT}_{\mu(X \cap A), i-1}[(X \cap \mathcal{V}) \cup A], \text{ and} \\ g_{v,i}(X) &= \text{OPT}_{v, i-1}[X \cup A]. \end{aligned}$$

Notice also that, if we interpret M to be an upper bound to the cost of any optimal solution of the original CLUBFS instance (e.g., by selecting $M = n^2$), then we have that $\text{OPT}_{v,i}[S \cup A]$, for every $i \geq |S|$ and for any $S \subseteq \mathcal{V}$, coincides with the cost of the optimal solution to the instance $\langle G', S, v \rangle$ of CLUBFS whenever such an instance is feasible. Otherwise, we have that $\text{OPT}_{v,i}[S \cup A]$ is at least M .

² The runtime originally given in Björklund et al. (2007) is here restated on our (implicitly assumed) model of computation, namely the standard unit-cost RAM with logarithmic word size, on which the $O(|X|^2 \cdot 2^{|X|})$ ring operations performed in Björklund et al. (2007) cost $O(W \cdot |X| \cdot \text{polylog}(W, |X|))$ time each. Notice that we are explicitly stating polynomial factors in $|X|$, i.e., logarithmic factors in $2^{|X|}$, which are disregarded in Björklund et al. (2007), since they will result in polynomial factors in k in the running time of our FPT algorithm.

Hence, the above relation can be exploited to define the following algorithmic process. We start by choosing $M = 1$ and then we perform a series of rounds as follows. In each round, we first determine all the values $\text{OPT}_{v,1}[H]$ by using Eq. (1). Then, for every $i = 2, \dots, k$, we compute n subset convolutions as shown in Eq. (3) (using Theorem 3). In more details, we compute $f_{v,i} * g_{v,i}$ of Eq. (3) for each vertex $v \in V$.

Finally, we set $\text{OPT}_{v,i}[H] = \min\{(f_{v,i} * g_{v,i})(H), M\}$ and we move to the next iteration. Here the minimum is necessary in order to ensure that the values computed by the subset convolutions that rely on $\text{OPT}_{v,i}[H]$ will be in $O(M)$. After the last iteration of this round is completed, $\text{OPT}_{s,k}[\mathcal{V} \cup A]$ stores either M or a value strictly smaller than M . On the one hand, if $\text{OPT}_{s,k}[\mathcal{V} \cup A] < M$, we have found the cost OPT of an optimal solution of the original instance, i.e., $\text{OPT} = \text{OPT}_{v,k}[\mathcal{V} \cup A]$. The optimal tree T_C^* can then be reconstructed from the values $\text{OPT}_{v,k}[S \cup A]$ for any $S \subseteq \mathcal{V}$, by using, e.g., the method in Dasgupta et al. (2008). On the other hand, if $\text{OPT}_{s,k}[\mathcal{V} \cup A] = M$, we move to the next round: we double the value of M and repeat the above procedure. We are now ready to give the following result.

Lemma 2 CLUBFS can be solved in $\tilde{O}(2^k k^3 n^4)$ time.

Proof First of all, notice that the cost of all the BFS trees of the clusters in $C \in \mathcal{V}$, from all the vertices $v \in V$, can be computed in $\tilde{O}(nm)$ time. Hence, it follows that all the $n \cdot 2^{|\mathcal{U}|} = n \cdot 2^{k+\log n} = n^2 \cdot 2^k$ base cases $\text{OPT}_{v,1}[H]$ can be computed in $O(n^3 + n^2 \cdot 2^k)$ time.

Now we focus on the values $\text{OPT}_{v,i}[H]$ having $i > 1$. In particular, notice that, for each M considered in the process, since functions $f_{v,i}$ and $g_{v,i}$ have values between 0 and $2M$, we can compute all n values $\text{OPT}_{v,i}[H]$ for each $H \subseteq U$, in

$$\begin{aligned} n \cdot \text{conv}(2M, U) &= n \cdot O(2M \cdot |U|^3 \cdot 2^{|\mathcal{U}|} \cdot \text{polylog}(2M, |U|)) \\ &= n \cdot O(M \cdot (k + \log n)^3 \cdot 2^{k+\log n} \cdot \text{polylog}(M, k + \log n)) \\ &= n \cdot O(M \cdot k^3 \cdot n \cdot 2^k \cdot \text{polylog}(n^2, n + \log n)) \\ &= \tilde{O}(M \cdot 2^k \cdot k^3 \cdot n^2). \end{aligned}$$

time. Overall, we perform at most $1 + \lceil \log \text{OPT} \rceil$ rounds, since we stop as soon as $M > \text{OPT}$ (i.e., when we have $M > \text{OPT}_{s,k}[\mathcal{V} \cup A] = \text{OPT}$). Hence, the overall time complexity of all rounds is

$$\begin{aligned} &= \tilde{O} \left(\sum_{M=1,2,4,\dots,2\text{OPT}} M \cdot 2^k \cdot k^3 \cdot n^2 \right) \\ &= \tilde{O}(\text{OPT} \cdot 2^k \cdot k^3 \cdot n^2) \\ &= \tilde{O}(2^k \cdot k^3 \cdot n^4), \end{aligned}$$

since $\Theta(n^2)$ is a trivial upper bound on the cost OPT of any feasible solution to CLUBFS. \square

It is worth noting that, in realistic settings, the number of clusters depends on various parameters, such as type of deployed devices and network density. However, it is almost always expected to be a small fraction w.r.t. overall number of vertices (see, e.g. Fareed et al. 2012; Sevgi and Kocyyigit 2008). Thus, CLUBFS–FPT1 might result in being truly effective in practice.

However, when this is not the case, then its running time might easily become impractical. In particular, if we focus on the classical BFS tree problem, which can be seen as a special instance of CLUBFS where each cluster contains only one vertex, it is easy to see that CLUBFS–FPT1 takes exponential time while the problem is known to be trivially solvable in $O(m + n)$ time! This suggests that, for the case in which \mathcal{V} consists of many singleton clusters, there must be another parametrization yielding a better complexity. Following this intuition, in the remaining of this section we present another FPT algorithm, namely CLUBFS–FPT2, parameterized in $h = |\{v \in V : v \in V_i, V_i \in \mathcal{V}, |V_i| > 1\}|$, i.e., in the total number of vertices that belong to clusters of size at least two.

2.2.2 Algorithm CLUBFS–FPT2

The idea underlying CLUBFS–FPT2 is as follows. Given a solution T to CLUBFS we call a *cluster root* for $V_i \in \mathcal{V}$ the unique vertex $v \in V_i$ with the smallest distance from s in T . The CLUBFS–FPT2 algorithm guesses the root of each cluster in an optimal solution T^* and then computes the optimal way of connecting the different roots of the clusters together.

Suppose we know a vector $\langle v_1, \dots, v_k \rangle$ of vertices such that $v_i \in V_i$. The key observation is that we can write the cost of any solution T having vertices v_1, \dots, v_k as cluster roots as follows:

$$\begin{aligned} \text{COST}(T) &= \sum_{V_i \in \mathcal{V}} \sum_{v \in V_i} d_T(s, v) \\ &= \sum_{V_i \in \mathcal{V}} \left(|V_i| d_T(s, v_i) + \sum_{v \in V_i} d_T(v_i, v) \right) \\ &= \sum_{V_i \in \mathcal{V}} |V_i| d_T(s, v_i) + \sum_{v \in V_i} d_T(v_i, v). \end{aligned}$$

Since $d_T(v_i, v) \geq d_{G[V_i]}(v_i, v)$, for any $v \in V_i$, the second summation is minimized when $d_T(v_i, v) = d_{G[V_i]}(v_i, v)$, i.e., when $T[V_i]$ is a BFS tree of $G[V_i]$. Consider now the first summation, and focus on its generic i th term. Let \mathcal{V}' be the set of clusters traversed by the path $\pi = \pi_T(s, v_i)$. For each cluster $V_j \in \mathcal{V}'$, let $x, y \in V_j$ be the first and last vertex of V_j traversed by π , respectively. By the definition of CLUBFS, and of cluster root, for V_i we have that: (i) all the vertices in the subpath of π between x and y , say $\pi[x, y]$, belong to V_i and (ii) $x = v_i$. Let P_i be the set of all the paths in G from s to v_i satisfying conditions (i) and (ii). It is easy to see that $d_T(s, v_i) \geq \min_{\pi' \in P_i} |\pi'|$. Hence, if T contains, for each $V_i \in \mathcal{V}$, the shortest path in P_i then $\sum_{V_i \in \mathcal{V}} |V_i| d_T(s, v_i)$ is minimized. To determine any path in P_i we proceed as follows. We define an auxiliary

directed graph G' , obtained from G by: (i) removing all the edges $(x, y) \in E$ such that neither x nor y is a root-vertex v_i for some i ; (ii) directing all the edges $(x, y) \in E$ such that x or y is a root-vertex v_i for some i towards v_i ; if both $x = v_i$ and $y = v_j$ (for some i, j) then we replace the undirected (x, y) by the pair of directed edges (x, y) and (y, x) ; (iii) replacing, for all $V_i \in \mathcal{V}$, all the edges in $E(G[V_i])$ with the edges of a BFS tree of $G[V_i]$ rooted in v_i . These edges are directed from the root towards the leaves of the tree. It is easy to see that any path in P_i is contained in G' , and that any BFS tree of G' must contain the edges of all the BFS trees of $G[V_i]$, hence minimizing $\text{COST}(T)$. Therefore the optimal solution to the instance of CLUBFS contains exactly the (undirected version of) the edges of a BFS tree of G' . The following lemma follows from the above discussion.

Lemma 3 CLUBFS–FPT2 solves CLUBFS in $O(h^{\frac{h}{2}}m)$ time.

Proof There are $\prod_{V_i \in \mathcal{V}} |V_i|$ ways of choosing a vector of cluster root vertices $\langle v_1, \dots, v_k \rangle$ for a given set \mathcal{V} of clusters. For each of these vectors the algorithm requires a computation of the BFS trees of $G[V_i]$ for $i = 1, \dots, k$ plus an additional BFS tree of G' . This can be done in $O(m) + \sum_{i=1}^k O(|E(G[V_i])| + |V_i|) = O(m)$ time. Finally, notice that $\prod_{i=1}^k |V_i| \leq h^{\frac{h}{2}}$ as the total number of clusters of size at least 2 is at most $\frac{h}{2}$. \square

Since it is possible to show that $T^*[V_i]$ must coincide with a BFS tree of $G[V_i]$ rooted at r_i , then this property allows us to efficiently reconstruct the optimal tree T^* , for a given guessed set of roots. Thus, overall, by combining CLUBFS–FPT1 and CLUBFS–FPT2, we can give the following result:

Theorem 4 CLUBFS can be solved in $\tilde{O}\left(\min\left\{2^k k^3 n^4, h^{\frac{h}{2}}m\right\}\right)$ time.

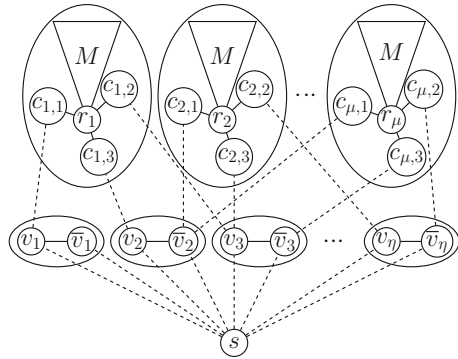
3 CLUSPT

In this section, we give our results on the CLUSPT problem. In particular, we first show that CLUSPT cannot be approximated, in polynomial time, within a factor of $n^{1-\epsilon}$ for any constant $\epsilon \in (0, 1]$, unless $P = NP$. Then, we give an n -approximation algorithm, thus proving that the mentioned inapproximability result is (essentially) tight. Finally, we show that, similarly to CLUBFS, CLUSPT is fixed-parameter tractable. Since CLUSPT is a generalization of CLUBFS, Theorem 1 immediately implies that CLUSPT is NP-hard as well. We can actually provide a stronger result, namely:

Theorem 5 CLUSPT cannot be approximated, in polynomial time, within a factor of $n^{1-\epsilon}$ for any constant $\epsilon \in (0, 1]$, unless $P = NP$.

Proof To prove the statement we use a slight modification of the construction given in the proof of Theorem 1. The main difference resides in the structure of the graph G_ϕ . In more details, for each clause c_j we do not add a triangle of vertices clustered into V_j . Instead, we add a subgraph to G_ϕ which is basically made of two components, as follows. First, we add four vertices, namely $c_{j,1}$, $c_{j,2}$, $c_{j,3}$ and r_j , to $V(G_\phi)$ and

Fig. 2 Graphical representation of the reduction used in the proof of Theorem 5



connect them in order to form a star graph with center r_j . Then, we create a tree of M vertices, where M is a parameter that will be specified later, which is connected to the above star graph through the center vertex r_j only. Finally, we cluster the two components together to form V_j . All edges have weight equal to zero, except those that connect the two vertices associated with a variable, which are unit-weighted. An example of the modified instance is shown in Fig. 2, where the triangle with label M represents a generic tree of M vertices, rooted, for each clause j , at vertex r_j . Now, by using an argument similar to that proposed in the proof of Theorem 1, it is easy to see that instance $\langle G_\phi, \mathcal{V}, s \rangle$, defined as above, exhibits the following properties: (i) if ϕ is satisfiable then $\text{OPT} = \eta$ (ii) if ϕ is not satisfiable then $\text{OPT} \geq \eta + M + 4$, where OPT denotes the cost of the optimal solution to the CLUSPT problem on instance $\langle G_\phi, \mathcal{V}, s \rangle$ and M can be chosen as an arbitrarily large integer.

We are now ready to prove the claim. Let $\langle G_\phi, \mathcal{V}, s \rangle$ be an instance of CLUSPT and let OPT be the cost of an optimal solution to such instance. Suppose by contradiction that there exists a polynomial-time $n^{1-\epsilon}$ -approximation algorithm A for CLUSPT for some constant $\epsilon \in (0, 1]$. Consider a 3-CNF-SAT instance along with the corresponding CLUSPT instance. W.l.o.g., let us assume that $\mu = \Theta(\eta)$. Note that NP-hard instances of 3-CNF-SAT of this latter kind are known to exist. We then set $M = \Theta(\eta^{2/\epsilon})$ so that the number of vertices of graph G_ϕ is $n = \Theta(\mu \cdot M) = \Theta(\eta^{1+2/\epsilon})$. If the 3-CNF-SAT instance is satisfiable, then A would return a solution T to the CLUSPT instance having a cost of at most: $\text{COST}(T) \leq n^{1-\epsilon} \eta = O(\eta^{\frac{2-2\epsilon}{\epsilon}} \eta) = O(\eta^{\frac{2}{\epsilon}-1}) = O(M\eta^{-1}) = o(M)$, while if it is not satisfiable $\text{COST}(T) \geq M$. Hence this would solve 3-CNF-SAT in polynomial time. \square

3.1 An approximation algorithm

We now show that the previous inapproximability result for CLUSPT is tight by providing a simple approximation algorithm, as stated in the following.

Theorem 6 *There exists a polynomial-time n -approximation algorithm for CLUSPT.*

Proof The algorithm works as follows: first it computes a multigraph G' from G by identifying each cluster $V_i \in \mathcal{V}$ into a single vertex v_i . When doing this, it associates

each edge of G' with the corresponding edge of G . Then it computes a *minimum spanning tree* (MST from now on) T' of G' , and k MSTs T_1, \dots, T_k of $G[V_1], \dots, G[V_k]$, respectively. Finally, the algorithm returns the spanning tree \tilde{T} of G which contains all the edges in $\bar{E} \cup \bigcup_{i=1}^k E(T_i)$, where \bar{E} denotes the set of edges of G associated with the edges of T' .

Let us now estimate the quality of \tilde{T} . Let T^* be an optimal solution to the CLUSPT instance. For a given spanning tree T of G rooted at s , let $w(T) = \sum_{e \in E(T)} w(e)$. Observe that clearly $w(T) \leq \text{COST}(T) \leq n \cdot w(T)$. Moreover, by construction, $w(\tilde{T}) \leq w(T^*)$. Thus, we have: $\text{COST}(\tilde{T}) \leq n \cdot w(\tilde{T}) \leq n \cdot w(T^*) \leq n \cdot \text{COST}(T^*)$. Since the time complexity is upper bounded by the complexity of computing the MSTs, the claim follows. \square

3.2 Fixed-parameter tractability results

The fixed-parameter tractability of CLUSPT directly follows from the discussion of Sect. 2 on the FPT algorithms for CLUBFS. In particular, if we focus on CLUBFS–FPT1, we observe that it can be trivially adapted to weighted graphs by considering SPTs instead of BFS trees, thus redefining the base cases $\text{OPT}_{v,1}[H]$ and the function $\ell(v, v')$. The only difference in the analysis is that it is no longer possible to use n^2 as an upper bound for the value of M . However, by retracing the calculations in the proof of Lemma 2, and by using the fact that $M = O(\text{OPT})$, one can easily prove the following:

Lemma 4 CLUSPT can be solved in $\tilde{O}(nm + 2^k k^3 n^2 \cdot \text{OPT} \log \text{OPT})$ time.

Regarding CLUBFS–FPT2, it can also be easily adapted to solve CLUSPT by using Dijkstra's algorithm instead of the BFS algorithm, when the solution to a sub-problem defined within each cluster has to be computed. This only slightly increases the resulting time complexity, which is however in the order of a logarithmic factor, as stated in the following.

Theorem 7 CLUSPT can be solved in $O\left(h^{\frac{h}{2}}(m + n \log n)\right)$ time.

Proof We prove the claim by elaborating on the proofs of Lemma 3. In particular, it suffices to note that in Algorithm CLUBFS–FPT2, for each vector of cluster root vertices, we need to compute the SPT trees (instead of BFS trees) of $G[V_i]$ for $i = 1, \dots, k$ plus an additional SPT tree of G' . This can be done in $O(m) + \sum_{i=1}^k O(|E(G[V_i])| + |V(G[V_i]) \log V(G[V_i])| + |V_i|) = O(m + n \log n)$ time. \square

To summarize, we can give the following theorem.

Theorem 8 CLUSPT can be solved in $\tilde{O}\left(\min\left\{nm + 2^k k^3 n^2 \cdot \text{OPT} \log \text{OPT}, h^{\frac{h}{2}} m\right\}\right)$ time.

4 CLUSP

To complement our results, we also studied CLUSP, i.e., the problem of computing a *clustered shortest path* between two given vertices of a graph. The problem was introduced in Lin and Wu (2016), and asks for finding a minimum-cost path, in a clustered *weighted* graph G , between a source and a destination vertex, with the constraint that in a feasible path, vertices belonging to a same cluster must induce a (connected) subpath. In this section, we extend the results of Lin and Wu (2016) by considering the unweighted version of the problem, which to the best of our knowledge was never considered before this work. We are then able to give the following result.

Theorem 9 *Unweighted CLUSP cannot be approximated, in polynomial time, within a factor of $n^{1-\epsilon}$ for any constant $\epsilon \in (0, 1]$, unless $P = NP$.*

Proof To prove the statement we show a polynomial-time reduction from the NP-complete problem Exact-Cover-by-3-Sets (X3C) to CLUSP. In the X3C problem we are given a set $\mathcal{I} = \{x_1, \dots, x_{3\eta}\}$ of 3η items, and a collection $\mathcal{S} = S_1, \dots, S_\mu$ of $\mu \geq \eta/3$ subsets of \mathcal{I} , each containing exactly 3 items. The problem consists of determining whether there exists a collection $\mathcal{S}^* \subset \mathcal{S}$ such that $|\mathcal{S}^*| = \eta$ and $\cup_{S \in \mathcal{S}^*} S = \mathcal{I}$ (i.e., each element of \mathcal{I} is contained in exactly one set of \mathcal{S}^*). For the sake of simplicity we assume that each $x_i \in \mathcal{I}$ is contained in at most 3 sets.³

Let M be an integer parameter that will be specified later. Given an instance $(\mathcal{I}, \mathcal{S})$ of X3C, the corresponding instance of CLUSP is constructed as follows:

- For each set $S_j \in \mathcal{S}$ we add four vertices u_j^0, u_j^1, u_j^2 , and u_j^3 .
- For $j = 1, \dots, \mu - 1$ we add the edge (u_j^3, u_{j+1}^0) .
- For each $x_i \in \mathcal{I}$ we add four vertices v_i^1, v_i^2, v_i^3 , and v_i . We connect v_i to each v_i^z , for $z = 1, 2, 3$, using a path of length M . The vertex v_i along with all the vertices in the paths from v_i to each v_i^z form a cluster.
- For each $S_j \in \mathcal{S}$ and $x_i \in S_j$, if x_i is the k th item in S_j , and S_j is the h th set to contain x_i , we add the edges (u_j^{k-1}, v_i^h) and (u_j^k, v_i^h) . For a given $S_j \in \mathcal{S}$, we call the set of the edges of the form (u_j^0, v_i^h) the *top-path* for S_j .
- For each $z = 1, \dots, \mu - \eta$, we add μ vertices y_z^1, \dots, y_z^μ and an additional vertex y_z connected to each y_z^1, \dots, y_z^μ with a path of length M . The vertex y_z along with all the vertices in the paths from y_z to each of y_z^1, \dots, y_z^μ form a cluster.
- For each set $S_j \in \mathcal{S}$ we add the $2(\mu - \eta)$ edges $\{(u_j^0, y_z^j) : z = 1, \dots, \mu - \eta\} \cup \{(u_j^3, y_z^j) : z = 1, \dots, \mu - \eta\}$. For a given $S_j \in \mathcal{S}$, we call the set of edges $\{(u_j^0, y_z^j), (u_j^3, y_z^j)\}$ the *z-th bottom-path* for S_j .

All vertices that have not explicitly been already assigned to a cluster belong to singleton clusters. Moreover we let $s = u_0^0$ and $t = u_\mu^3$. An example of the above construction is shown in Fig. 3.

Now, let OPT be the cost of an optimal solution to this CLUSP instance. We now claim that (i) if there is a solution for the X3C instance then $\text{OPT} \leq 15\mu$, and (ii)

³ The X3C problem remains NP-complete even with this additional assumption, see e.g., problem SP2 in Garey and Johnson (1979).

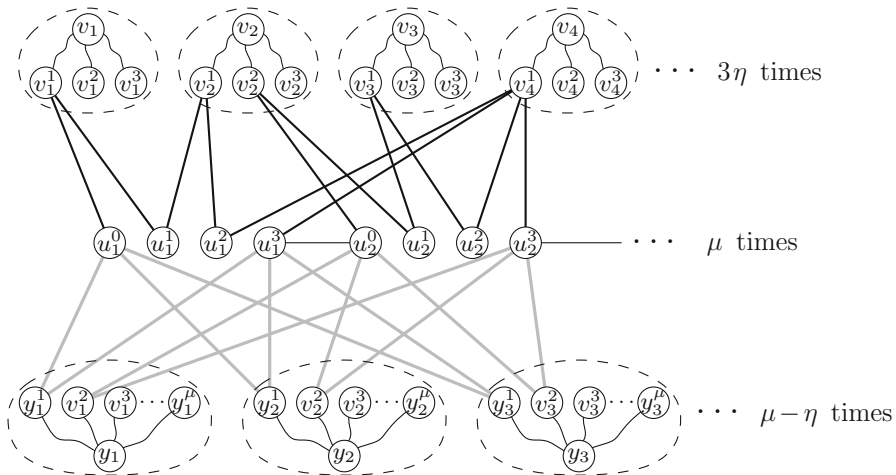


Fig. 3 The graph used in the proof of Theorem 9. We show the reduction for the first two sets of an instance of X3C, where we assume that $S_1 = \{x_1, x_2, x_4\}$ and $S_2 = \{x_2, x_3, x_4\}$. Top-paths are shown with bold black edges. Bottom paths are shown with bold-gray edges. Paths of length M are shown with curvy lines. Clusters are shown with dashed lines. In the corresponding X3C instance we have $S_1 = \{x_1, x_2, x_4\}$ and $S_2 = \{x_2, x_3, x_4\}$

if there is no solution to the X3C instance then $\text{OPT} \geq M$. To prove (i), let \mathcal{S}^* be a solution to the X3C instance. Notice that $|\mathcal{S}^*| = \eta$ and that $|\mathcal{S} \setminus \mathcal{S}^*| = \mu - \eta$. We construct a clustered s - t -path P as follows: for each $S_j \in \mathcal{S}$, if $S_j \in \mathcal{S}^*$ we add to P all the edges in the top-path for S_j , while if $S_j \notin \mathcal{S}^*$ we let $z = |\{S_1, \dots, S_j\} \setminus \mathcal{S}^*|$ and we add to P all the edges in the z th bottom-path for S_j . Finally, we add to P all the edges in $\{(u_j^3, u_{j+1}^0) : j = 1, \dots, \mu - 1\}$. It is easy to see that P is indeed an s - t -path and that each cluster is traversed only once. Moreover, P contains exactly η top-paths (of 6 edges each) and $\mu - \eta$ bottom paths (of 2 edges each). Therefore, it follows that: $\text{OPT} \leq 6\eta + 2(\mu - \eta) + \mu - 1 \leq 4\eta + 3\mu \leq 15\mu$.

To prove (ii) we consider the contrapositive statement, i.e., we show that if $\text{OPT} < M$ then there exists a solution to the X3C instance. Let P^* be an optimal solution to the CLUSP instance and suppose $\text{OPT} < M$. This immediately implies that P does not contain any of the paths from y_z to y_z^j or any of those from v_i to v_i^h , since all these paths have length M . This means that, for each $S_j \in \mathcal{S}$, P contains either the (unique) top-path for S_j or one of the bottom paths for S_j . Since P can contain at most $\mu - \eta$ bottom-paths and at most η top paths (as otherwise it would violate the clustering constraints), it follows that P contains *exactly* η top paths. We define \mathcal{S}^* as the collection of the sets S_j for which a top-path has been selected. Since $|\mathcal{S}^*| = \mu$ and two paths corresponding to two different sets in \mathcal{S}^* cannot both pass through vertices belonging to the same cluster, it follows that \mathcal{S}^* is indeed a solution to the X3C instance.

We are now ready to prove the claim. Notice that the number of vertices of the CLUSP instance, say n , is upper bounded by $O(\mu^2 M)$. We set $M = \Theta(\mu^{\frac{3}{\epsilon}-1})$ so that $n = O(\mu^{\frac{3}{\epsilon}+1})$. Suppose now that there exists a polynomial-time $n^{1-\epsilon}$ -approximation

algorithm A for CLUSP. This would imply that if the X3C instance admits a solution, then the cost of the solution returned by A would be at most:

$$15\mu n^{1-\epsilon} = O(\mu\mu^{\frac{3}{\epsilon}-2-\epsilon}) = O(\mu^{\frac{3}{\epsilon}-1-\epsilon}) = O(M\mu^{-\epsilon}) = O(MM^{-\frac{\epsilon^2}{3-\epsilon}}) = o(M)$$

while, if the X3C instance does not admit a solution, then A would return a solution to the CLUSP instance having a cost of at least M . It follows that we would be able to solve X3C in polynomial time. \square

5 Conclusion and future work

In this paper, motivated by key modern networked applications, we have studied several clustered variants of shortest-path related problems, namely CLUBFS, CLUSPT and unweighted CLUSP. We have provided a comprehensive set of results which allow to shed light on the complexity of such problems.

There are several directions that may be pursued for future work. The main research question that we leave open is that of establishing a lower bound on the approximability of CLUBFS, and, in case of a gap w.r.t. the approximation factor provided by Algorithm 1, that of devising a better approximation algorithm (by, e.g., exploring some other natural heuristic). Besides that, also studying clustered shortest-path problems on restricted but meaningful classes of graphs, like, e.g., euclidean or planar graphs, might deserve investigation. Another interesting issue is surely that of studying how other practically relevant network structures, such as spanners and highly-connected spanning subgraphs, behave in a clustered setting [incidentally, clusterization is one of the most used techniques to build this kind of structures, see e.g. Bilò et al. (2015)]. Finally, it would be also interesting to conduct an experimental study for assessing the practical performance of all proposed algorithms.

References

- Bao X, Liu Z (2012) An improved approximation algorithm for the clustered traveling salesman problem. *Inf Process Lett* 112(23):908–910
- Bilò D, Grandoni F, Gualà L, Leucci S, Proietti G (2015) Improved purely additive fault-tolerant spanners. In: *Proceedings 23rd European symposium on algorithms (ESA)*, volume 9294 of *Lecture notes in computer science*. Springer, pp 167–178
- Björklund A, Husfeldt T, Kaski P, Koivisto M (2007) Fourier meets möbius: fast subset convolution. In: *Proceedings 39th ACM symposium on theory of computing (STOC)*. ACM, pp 67–74
- Byrka J, Grandoni F, Rothvoß T, Sanità L (2013) Steiner tree approximation via iterative randomized rounding. *J ACM* 60(1):6
- Dasgupta S, Papadimitriou CH, Vazirani U (2008) *Algorithms*, 1st edn. McGraw-Hill Inc, New York
- D’Emidio M, Forlizzi L, Frigioni D, Leucci S, Proietti G (2016) On the clustered shortest-path tree problem. In: *Proceedings 17th Italian conference on theoretical computer science (ICTCS)*, volume 1720 of *CEUR workshop proceedings*, pp 263–268
- Fareed MS, Javaid N, Akbar M, Rehman S, Qasim U, Khan ZA (2012) Optimal number of cluster head selection for efficient distribution of sources in WSNs. In: *Proceedings seventh international conference on broadband, wireless computing, communication and applications*. IEEE, pp 626–631
- Feremans C, Labbé M, Laporte G (2003) Generalized network design problems. *Eur J Oper Res* 148(1):1–13

- Garey MR, Johnson DS (1979) *Computers and intractability: a guide to the theory of NP-completeness*. W. H. Freeman & Co, New York
- Garg N, Konjevod G, Ravi R (2000) A polylogarithmic approximation algorithm for the group Steiner tree problem. *J Algorithms* 37(1):66–84
- Guttmann-Beck N, Hassin R, Khuller S, Raghavachari B (2000) Approximation algorithms with bounded performance guarantees for the clustered traveling salesman problem. *Algorithmica* 28(4):422–437
- Halperin E, Krauthgamer R (2003) Polylogarithmic inapproximability. In: *Proceedings 35th ACM symposium on theory of computing (STOC)*, pp 585–594
- Lin C, Wu BY (2016) On the minimum routing cost clustered tree problem. *J Comb Optim* 31(1):1–16
- Sevgi C, Kocyyigit A (2008) On determining cluster size of randomly deployed heterogeneous WSNs. *IEEE Commun Lett* 12(4):232–234
- Wu BY, Lancia G, Bafna V, Chao K-M, Ravi R, Tang CY (1998) A polynomial time approximation scheme for minimum routing cost spanning trees. In: *Proceedings 9th ACM-SIAM symposium on discrete algorithms (SODA)*, pp 21–32
- Wu BY, Lin C (2015) On the clustered Steiner tree problem. *J Comb Optim* 30(2):370–386
- Zou P, Li H, Wang W, Xin C, Zhu B (2018) Finding disjoint dense clubs in a social network. *Theor Comput Sci* 734:15–23

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Mattia D'Emidio¹  · Luca Forlizzi¹ · Daniele Frigioni¹ · Stefano Leucci² · Guido Proietti^{1,3}

Luca Forlizzi
luca.forlizzi@univaq.it

Daniele Frigioni
daniele.frigioni@univaq.it

Stefano Leucci
stefano.leucci@inf.ethz.ch

Guido Proietti
guido.proietti@univaq.it

¹ Department of Information Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy

² Department of Computer Science, ETH Zürich, Zürich, Switzerland

³ Istituto di Analisi dei Sistemi e Informatica "Antonio Ruberti" Consiglio Nazionale delle Ricerche, Rome, Italy