International Journal of Foundations of Computer Science
Vol. 34, No. 5 (2023) 453–468

DOI: 10.1142/S0129054122500265

World Scientific
www.worldscientific.com

Online and Approximate Network Construction from Bounded Connectivity Constraints*

Jesper Jansson

Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
jj@i.kyoto-u.ac.jp

Christos Levcopoulos[†] and Andrzej Lingas[‡,§]

Department of Computer Science, Lund University
Box 118, 22100 Lund, Sweden
[†]*Christos.Levcopoulos@cs.lth.se*
[‡]*Andrzej.Lingas@cs.lth.se*

Received 29 April 2022
Accepted 5 November 2022
Published 6 January 2023
Communicated by Oscar Ibarra

The *Network Construction problem*, studied by Angluin *et al.*, Hosoda *et al.*, and others, asks for a minimum-cost network satisfying a set of *connectivity constraints* which specify subsets of the vertices in the network that have to form connected subgraphs. More formally, given a set V of vertices, construction costs for all possible edges between pairs of vertices from V, and a sequence $S_1, S_2, \ldots, S_r \subseteq V$ of connectivity constraints, the objective is to find a set E of edges such that each S_i induces a connected subgraph of the graph (V, E) and the total cost of E is minimized. First, we study the online version where every constraint must be satisfied immediately after its arrival and edges that have already been added can never be removed. We give an $O(B^2 \log n)$-competitive and $O((B+\log r)\log n)$-competitive polynomial-time algorithms, where B is an upper bound on the size of constraints, while r, n denote the number of constraints and the number of vertices, respectively. On the other hand, we observe that an $\Omega(B)$-competitive lower bound as well as an $\Omega(\sqrt{B})$-competitive lower bound in the cost-uniform case are implied by the known lower bounds for unbounded constraints. For the cost-uniform case with unbounded constraints, we provide an $O(\sqrt{n}(\log n + \log r))$-competitive upper bound with high probability. The latter bound is against an oblivious adversary while our other randomized competitive bounds are against an adaptive adversary. Next, we discuss a hybrid approximation method for the (offline) Network Construction problem combining an approximation algorithm of Hosoda *et al.* with one of Angluin *et al.* and an application of the hybrid method to bioinformatics. Finally, we consider a natural strengthening

*This article is an extended version of the paper that appeared in Proceedings of the 12th International Conference on Algorithms and Complexity (CIAC 2021), *Lecture Notes in Computer Science*, volume 12701, pp. 314–325, 2021.

[§]Corresponding author.

of the connectivity requirements in the Network Construction problem, where each constraint has to induce a subgraph (of the constructed graph) of diameter at most d. Among other things, we provide a polynomial-time $(\binom{B}{2} - B + 2)\binom{B}{2}$-approximation algorithm for the Network Construction problem with the d-diameter requirements, when each constraint has at most B vertices, and show the APX-completeness of this variant.

Keywords: Network optimization; induced subgraph; connectivity; approximation algorithm; online algorithm.

1. Introduction

Korach and Stern introduced the problem of interconnecting possibly overlapping groups of users by a network such that the users in the same group do not need to use connections outside the group [13]. The optimization objective is to minimize the total cost of the pairwise connections. Angluin *et al.* and Chockler *et al.* studied this problem in [2] and [6], respectively.

Angluin *et al.* showed in [2] that if $P \neq NP$ and n is the number of vertices in the network then the problem cannot be approximated within a factor that is sublogarithmic in n, even in the uniform edge cost case. On the other hand, they proved that a greedy heuristic can approximate the optimal solution within a factor of $O(\log r)$, where r is the number of constraints. As observed in [2], the lower bound matches the upper bound in case r is polynomial in n.

Angluin *et al.* also studied the online version of this problem where each constraint has to be satisfied directly after its arrival and edges that have already been added can never be removed [2]. Their motivation for this problem variant was to help infer the structure of a social network describing the spread of diseases in a community and to decide where to allocate resources to fight an epidemic efficiently. They assumed that the individuals affected by each outbreak of a disease are specified by a connectivity constraint, that the outbreaks occur over time, and that resources that have been committed cannot be released. They provided an $O(n \log n)$-competitive online algorithm for the online version along with an $\Omega(n)$-competitive lower bound. They also considered the uniform cost case of this online version, providing an $O(n^{2/3} \log^{2/3} n)$-competitive algorithm against an oblivious adversary and an $\Omega(\sqrt{n})$-competitive lower bound against an adaptive adversary.

Hosoda *et al.* studied a *B-constraint-bounded* variant of the Network Construction problem, where the cardinality of each connectivity constraint S_i does not exceed B [10]. This corresponds to constructing a minimum overlay network for a topic-based peer-to-peer pub/sub system where users (represented by vertices) who are interested in a common topic (represented by connectivity constraints) form connected subgraphs, and moreover, the number of users following each topic is bounded by a constant due to the publisher of that topic having a limited number of available slots for users. Hosoda *et al.* provided a polynomial-time $(\lfloor B/2 \rfloor \lceil B/2 \rceil)$-approximation algorithm for this variant and proved its APX-completeness in [10].

A natural generalization of the Network Construction problem, where some pairwise connections are given a priori has applications in bioinformatics [15]. The purpose is to infer protein-protein interactions that are missing from a database based on a collection of known, overlapping protein complexes (see Sec. 4.1 for details).

1.1. *Article outline and contributions*

The next section defines the Network Construction problem and its B-constraint-bounded variant, where each constraint includes at most B vertices. We also review the Minimum Weight Set Cover problem and some facts about its approximability. In Sec. 3, we study the online version of the B-constraint-bounded Network Construction problem. We present $O(B^2 \log n)$-competitive and $O((B+\log r)\log n)$-competitive polynomial-time algorithms for the online B-constraint-bounded Network Construction problem, where r, n stand for the number of constraints and the number of vertices in the network, respectively. In the cost-uniform case when constraints are unbounded, we provide an $O(\sqrt{n}(\log n + \log r))$-competitive upper bound with high probability. The latter bound is against an oblivious adversary while our other randomized competitive bounds are against an adaptive adversary. We also observe that a $(B-1)$-competitive lower bound in case of arbitrary edge costs and a $\sqrt{B}$-competitive lower bound in case of uniform edge costs are implied by the aforementioned lower bounds of Angluin *et al.* for unbounded constraints [2]. In Sec. 5, we study approximation algorithms for the offline Network Construction problem and its extensions. First, we discuss a hybrid approximation method combining the approximation algorithm of Hosoda *et al.* from [10] with that of Angluin *et al.* from [2] in the context of the application to bioinformatics. Next, we consider a natural strengthening of the connectivity requirements in the Network Construction problem. Each constraint has to induce a subgraph (of the constructed graph) of diameter at most d, where d is given a priori. We provide a polynomial-time $(\binom{B}{2} - B + 2)\binom{B}{2}$-approximation algorithm for the aforementioned problem with the d-diameter requirements, when each constraint has at most B vertices. Also, we show the restricted variant to be APX-complete. Furthermore, we present a polynomial-time algorithm achieving a non-trivial approximation ratio in the general case of the d-diameter variant, where the cardinality of constraints is unbounded. We conclude with final remarks.

Our approximate or online solutions to the aforementioned variants with bounded constraints can be used to solve approximately or online the corresponding variants with unbounded constraints by splitting the constraints into small and large ones (Secs. 4 and 5).

2. Preliminaries

For a positive integer r, the term $[r]$ will denote $\{1, \dots, r\}$, and for sets S and V, $|S|$ will stand for the cardinality of S while V^2 for $\{\{v, u\}|v, u \in V\}$.

We shall consider simple graphs (i.e., graphs without loops and multiple edges). A *subgraph* of a graph (V, E) is a graph (V', E') such that $V' \subseteq V$ and $E' \subseteq E$. The

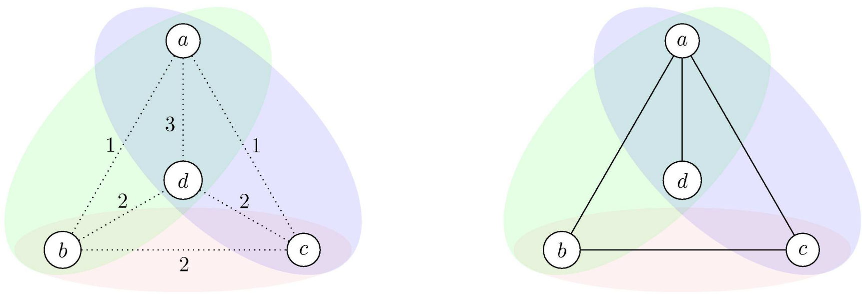

Fig. 1. An instance of the Network Construction problem with $V = \{a, b, c, d\}$, $S = \{\{a, b, d\}, \{a, c, d\}, \{b, c\}\}$, and edge costs as indicated on the left. The optimal solution (total cost: 7) is shown on the right. Note that the optimal solution does not induce a locally optimal solution for each connectivity constraint, e.g., $\{(a, b), (a, d)\}$ is not an optimal solution for $\{a, b, d\}$. Also note that in this example, the optimal solution is not a tree.

subgraph of a graph (V, E) *induced* by a subset S of V is the graph $(S, E \cap S^2)$. A *perfect cut* of a graph (V, E) is a partition of V into subsets V' and V'' such that $E \cap \{\{v, u\} | v \in V' \ \& \ u \in V''\} = \emptyset$. The *diameter* of a graph (V, E) is the minimum number ℓ such that any pair of vertices in V can be connected by a path composed of at most ℓ edges in E. If the graph is disconnected, its diameter is undefined.

The *Network Construction* problem is as follows [2]. We are given a set V of vertices and for each possible edge $e = \{v_i, v_j\}$, the nonnegative real cost $c(e)$ of its construction. We are also given a collection of connectivity constraints $S = \{S_1, \ldots, S_r\}$, where each S_i is a subset of V. The objective is to construct a set E of edges in V^2 such that for $i = 1, \ldots, r$, the subgraph of the graph (V, E) induced by S_i is connected and the total cost of the edges in E is minimized. See Fig. 1 for an example. In the *uniform-cost* case of the problem, we have $c(e) = 1$ for all edges in V^2. We can naturally generalize the problem to include the *Network Extension* problem, where some subset E' of edges is already given (constructed) a priori. Note that when zero construction costs of edges are allowed the Network Construction problem is equivalent to the Network Extension problem. Simply it is sufficient to set the construction costs of the edges given a priori to zero in order to obtain an equivalent version of the Network Construction problem. In order to avoid duplications in our statements, in the aforementioned situation we shall mention only the Network Construction problem.

Among other things, the following fact was established by Angluin *et al.* in [2].

Fact 1 ([2, Theorem 2]). *There is a polynomial-time $O(\log r)$-approximation algorithm for the Network Construction problem on r constraints.*

We shall also consider a *B-constraint-bounded* variant of the Network Construction problem, where the cardinality of each connectivity constraint S_i does not exceed B. It was studied by Hosoda *et al.* in [10]. They provided a polynomial-time approximation algorithm for this variant and showed its APX-completeness.

Fact 2 ([10, Theorem 4]). *The B-constraint-bounded Network Construction problem admits a polynomial-time $\lfloor B/2 \rfloor \lceil B/2 \rceil$-approximation (i.e., $\approx B^2/4$-approximation) algorithm.*

In the context of the Network Construction and Extension problems, we refer to two types of edges: those already constructed and the remaining ones that potentially could be constructed. For instance, when referring to a perfect cut, we consider the edges of the first type while when we refer to edges crossing a perfect cut we mean the edges of the second type.

Recall the definition of *Minimum Weight Set Cover* problem. The input to this problem is a universal set U on n elements and a family F of m subsets of U. Each subset in F is assigned a non-negative weight. A *set cover* is a sub-family of F whose union is equal to U. The objective is to find a set cover of minimum total weight. The decision version of this optimization problem is already NP-hard in the uniform-weight case [8]. The so-called *Minimum Weight Hitting Set* problem is an equivalent formulation of the Minimum Weight Set Cover problem with the roles of elements and subsets exchanged. Here, the input is a finite set S of weighted elements and a family C of subsets of S. The objective is to find a minimum weight subset of S that hits all the subsets in C, i.e., that has a non-empty intersection with each of the subsets in C. This problem is known to be equivalent to Minimum Weight Set Cover [3]. Consequently, approximation algorithms and inapproximability results for each of them carry over to the other one.

Hochbaum [9] used a relaxation of an integer linear programming formulation to obtain an approximation of the Minimum Weight Set Cover in cubic time. The same approximation ratio was obtained by Bar-Yehuda and Even [4] with a more direct, linear-time method. We summarize their results as follows.

Fact 3. *The Minimum Weight Set Cover problem (U, F), where each element of the universal set U occurs in at most B subsets of U in F, can be approximated within the multiplicative factor B in linear time. Consequently, the Minimum Weight Hitting set problem, where each subset in the given family has cardinality at most B, can be approximated within B in linear time.*

A decade later, Papadimitriou and Yannakakis showed the following fact [16].

Fact 4. *The Minimum Weight Set Cover problem (U, F), where each element of the universal set U occurs in at most $B \geq 2$ subsets of U in F, is APX-complete (already in the uniform weight case). Consequently, the Minimum Weight Hitting Set problem, where each subset in the given family has cardinality at most B, is APX-complete (already in the uniform weight case).*

The Minimum Weight Hitting Set problem, where each subset in the given family has cardinality at most B, is easily seen to be equivalent to Vertex Cover in B-uniform hypergraph. More recently, the latter problem has been shown to be NP-hard to approximate beyond $(B-1)$ factor by Dinur *et al.* [7] and UGC-hard to approximate beyond B factor by Khot and Regev [12]. (For the Unique Games Conjecture and the related concept of UGC-hardness see [11]). Hence, we have the following fact.

Fact 5. *The Minimum Weight Hitting Set problem, where each subset in the given family has cardinality at most B, is NP-hard to approximate beyond $(B-1)$ factor* [7] *and UGC-hard to approximate beyond B* [12].

3. Online B-Constraint-Bounded Network Construction

In this section, we consider the online version of the Network Construction problem studied in [2]. It arises naturally in the situation when the knowledge about the relationships between the entities represented by vertices changes over time. In the online version, the vertices, potential edges and their costs are given beforehand and the collection of connectivity constraints is given one at a time. When a constraint S_i is presented, the online algorithm is in round i. The algorithm now has to satisfy this constraint during this round by constructing, if necessary, additional edges before the start of the next round (no previously constructed edges may be removed). The next constraint is then presented in round $i+1$. To study the worst-case performance of our online algorithms, we shall use an *adaptive adversary* that can wait with setting the next constraint until the online algorithm satisfied the previous one. We shall use competitive analysis of our online algorithms. The idea is to compare an online algorithm ALG to an optimal offline algorithm OPT that knows the entire request sequence in advance. For a request sequence δ, let $ALG(\delta)$ and $OPT(\delta)$ denote the costs incurred by ALG and OPT, respectively. Algorithm ALG is called *c-competitive* if there exists a constant b such that $ALG(\delta) \leq cOPT(\delta) + b$, for all sequences δ. c is called the *competitive ratio* of algorithm ALG.

3.1. *Upper bounds*

First consider the following online Fractional Network Construction problem: for a set V of vertices and edge costs $c(e)$ for $e \in V^2$, and sequence of connectivity constraints $S_1, \ldots, S_r$, assign fractional capacities $w(e)$ to the edges e such that for each $i \in [r]$, for each pair of vertices in S_i, the maximum flow between them is at least 1. More precisely, after designating one of the vertices as source and the other as sink by flow we mean here an assignment of fractional flow values to the edges e in S_i^2 that do not exceed $w(e)$ and obey the flow preservation rule, i.e., the total incoming flow is equal to the total outgoing flow. The optimization objective is to minimize $\sum_e w(e)c(e)$.

Fact 6 ([2, Lemma 2]). *There is an $O(\log n)$-competitive polynomial-time algorithm for the online Fractional Network Construction problem on n vertices.*

By using this fact, we obtain the following theorem.

Theorem 1. *There is an $O(B^2 \log n)$-competitive polynomial-time algorithm for the online B-constraint-bounded Network Construction problem on n vertices.*

Proof. Run the online $O(\log n)$-competitive algorithm for the online Fractional Network Construction problem from Fact 6. Disregard all edges that are assigned capacity smaller than B^{-2} by the online solution to the fractional problem and construct all the remaining edges. Note that after the edges of capacity smaller than B^{-2} have been removed, the maximum flow between each pair of vertices in any B-bounded constraint is still at least $1 - \binom{B}{2} B^{-2} \geq \frac{1}{2}$. Hence, there is a path composed of the constructed edges between such a pair. The cost of the constructed edges is at most B^2 times larger than the cost of the fractional solution, i.e., the sum of products of edge cost and edge capacity over all edges. □

To derive another competitive upper bound for the online B-constraint-bounded Network Construction problem, we shall consider the online version of the Minimum Weight Set Cover problem. In this version, a family of subsets of the universal set is given beforehand (and hence we know the elements forming the subsets) but then the elements of a subset of the universal set are presented online one at a time [1]. A new element has to be covered before the arrival of the next one. Analogously, in the online version of the equivalent Minimum Weight Hitting set problem, the set of hitting elements is given a priori, and the sets to be hit arrive online one at a time. A new set has to be hit before the arrival of the next one. Alon *et al.* established the following fact in [1].

Fact 7. *There is an $O(\log n \log r)$-competitive polynomial-time algorithm for the online Minimum Weight Set Cover problem, where n is the cardinality of the universal set and r is the cardinality of the given family of subsets of the universal set. Consequently, there is an $O(\log n \log r)$-competitive polynomial-time algorithm for the online Minimum Weight Hitting set, where n is the cardinality of the family of sets to hit and r is the cardinality of the set of all possible hitting elements.*

By combining Fact 7 with the reduction of the Network Construction problem to the Minimum Weight Set Cover problem given by Angluin *et al.* in [2], we obtain another competitive upper bound for the online B-constraint-bounded Network Construction problem.

Theorem 2. *There is an $O((B + \log r) \log n)$-competitive polynomial-time algorithm for the online B-constraint-bounded Network Construction problem with n vertices and r constraints.*

Proof. We shall reduce the online Network Construction problem to the online Minimum Weight Hitting Set problem, following the reduction of the former problem to the online Minimum Weight Set Cover problem from [2]. The set of the possible hitting elements given a priori is just the set of all possible edges. Each edge has weight equal to the cost of its construction. For a constraint S_i, let $PC(S_i)$ be the set of perfect cuts induced by the constraint upon its arrival online. Next, for each perfect cut $C \in PC(S_i)$, let $AE(C)$ be the set of all (additional potential) edges crossing the perfect cut, i.e., having endpoints in the two different parts of the bipartition. Note that the constraint S_i is satisfied if for each perfect cut $C \in PC(S_i)$ there is an edge in $AE(C)$ accounted to the online formed hitting set. It follows that the cost of an optimal solution to the resulting online Minimum Hitting Set problem is the same as that to the original Network Construction problem. Now it is sufficient to observe that the former problem has $O(n^2)$ possible hitting elements and at most $r2^B$ sets to hit, and then to apply Fact 7. $\square$

We can use Theorem 2 to derive a competitive upper bound for the uniform-cost variant of the Network Construction problem with unbounded constraints. Angluin *et al.* considered also the uniform cost variant of the Network Construction problem in [2], providing an $O(n^{2/3} \log^{2/3} n)$-competitive algorithm against an oblivious adversary and an $\Omega(\sqrt{n})$-competitive lower bound against an adaptive adversary. Our upper bound in the uniform case is as that of Angluin *et al.* against an oblivious adversary, i.e., an adversary not knowing the randomized results of the algorithm. The key idea is to split the constraints into small and large ones, and use Theorem 2 to process the former ones.

Theorem 3. *The uniform cost online Network Construction problem on* n *vertices and* r *constraints admits an* $O(kn^{0.5}(\log n + \log r))$*-competitive polynomial-time algorithm, for every positive integer* k*, with probability at least* $1 - O((nr)^{-1}) - \frac{1}{k}$ *provided that* n, r *are known in advance and* $\ln n + 2\ln r < n^{0.5}$*, any feasible solution is connected, and the adversary is oblivious.*

Proof. Split the set of constraints into two sets, one consisting of all constraints of size $\leq n^{0.5}$ and one consisting of the rest. We can apply the $O((B + \log r)\log n)$ competitive algorithm from Theorem 2 to the small constraints obtaining an $O((n^{0.5} + \log r)\log n)$ competitive solution. To satisfy the large constraints with more than $n^{0.5}$ vertices we proceed as follows.

We initialize an empty vertex set W. Upon the arrival of a new large constraint, each vertex v in the constraint that for the first time occurs in a large constraint is added to W with probability $n^{-0.5}(\ln n + 2\ln r)$. We have $\ln n + 2\ln r < n^{0.5}$ by the theorem assumptions. Furthermore, if v is added to W then all missing (i.e., not yet constructed) edges incident to v are constructed. It follows that the expected total number, and hence, the expected total cost of the so constructed edges amounts to at most $n^{0.5}(n-1)(\ln n + 2\ln r)$. Thus, for a positive integer parameter k, the total cost is at most $kn^{0.5}(n-1)(\ln n + 2\ln r)$ with probability at least $1 - \frac{1}{k}$ by Markov's inequality. Since the adversary is oblivious, for each large constraint the probability that it does not contain any vertex from W (before the next constraint arrival, if any; otherwise, at the end of the execution) is at most

$$\left(1 - \frac{\ln n + 2\ln r}{n^{0.5}}\right)^{n^{0.5}} \leq e^{-\ln n - 2\ln r} \in O\left(\frac{1}{nr^2}\right).$$

Since there are at most r large constraints and the cost of an optimal solution is at least $n-1$, we obtain an $kn^{0.5}(\ln n + 2\ln r)$ competitive upper bound for the large constraints with probability at least $1 - O((nr)^{-1}) - \frac{1}{k}$. □

The choice of the threshold $n^{0.5}$ in order to split the constraints into small and large ones in the proof of Theorem 3 seems to be asymptotically optimal as long as r is not substantially larger than n. Otherwise, multiplying $n^{0.5}$ by a slowly growing function of r yields a better upper bound on the competitive ratio, e.g., $n^{0.5}\sqrt{\log r}$ yields $O(kn^{0.5}\sqrt{\log r}\log n)$ by $r \leq 2^n$.

3.2. *Lower bounds*

We present two lower bounds on the competitiveness of algorithms for the online B-constraint-bounded Network Construction problem implied by the corresponding general lower bounds established in [2].

When the edge costs can be arbitrary, we can use the following fact due to Angluin *et al.* [2].

Fact 8. *There is no (cn)-competitive algorithm for $c < 1$ for the online Network Construction problem on n vertices, even when the underlying graph is a path.*

By setting $n = B$, we obtain immediately the following corollary from Fact 8.

Corollary 4. *For any $c < 1$, there is no cB-competitive algorithm for the online B-constraint-bounded Network Construction problem.*

For the uniform cost case, we can use the following fact due to Angluin *et al.* [2].

Fact 9. *The online uniform cost Network Construction problem on n vertices has an $\Omega(\sqrt{n})$-competitive lower bound.*

Again, by setting $n = B$, we obtain immediately the following corollary from Fact 9.

Corollary 5. *The online uniform cost B-constraint-bounded Network Construction problem has an $\Omega(\sqrt{B})$ competitive lower bound.*

4. Offline Approximation Algorithms

In this section, we first discuss a hybrid approximation method for the offline Network Construction problem and its application to bioinformatics. It combines the approximation algorithm of Hosoda *et al.* from [10] with that of Angluin *et al.*

from [2]. Next, we present approximation algorithms for a strengthened version of the Network Construction problem, where each constraint has to induce a subgraph (of the constructed graph) of diameter at most d for d given a priori.

4.1. *A hybrid method with biological applications*

An application of the Network Extension problem to bioinformatics was given in [15]. There, the goal was to infer protein-protein interactions (PPIs) that were missing from a database based on a collection of known, overlapping protein complexes. More precisely, the vertices V in the input graph were used to represent proteins, the set E' of a priori given edges represented PPIs already in the database, and each input connectivity constraint S_i consisted of the proteins belonging to a single protein complex. Using the assumption that each protein complex must induce a connected subgraph, solving instances of the Network Extension problem gave lower bounds on the number of missing PPIs in various widely used PPI databases. The overwhelming majority of complexes in the existing PPI databases seem to be of small size, containing at most 10 proteins each, but a few larger ones with up to 100 proteins also occur (for details, see Table A3 in the Supporting Information file for [15]).

The aforementioned statistics suggest a hybrid method consisting of applying the approximation algorithm of Hosoda *et al.* from [10] to the constraints corresponding to small complexes and that of Angluin *et al.* from [2] to the constraints corresponding to larger complexes. We can express it in terms of the Network Construction problem by the equivalence observed in Sec. 2. The output is the union of the output of each of the two algorithms applied separately. Hence, by combining Fact 1 with Fact 2, we obtain the following theorem.

Theorem 6. *Consider an instance of the Network Construction problem. For* $B \in [n]\setminus\{1\}$, *let* r_B *be the number of constraints with more than* B *vertices in the instance. A solution to the instance (for the respective problem) of total cost not exceeding* $\min_{B\in[n]\setminus\{1\}} \lfloor B/2 \rfloor \lceil B/2 \rceil + O(\log r_B)$ *times the minimum can be found in polynomial time.*

The hybrid method will be useful when there is a relatively small $B \in [n]\setminus\{1\}$ such that the number r_B of large constraints including more than B vertices, i.e., the number of large complexes in the biological application, is small.

Once B is chosen, one can refine the hybrid approximation method as follows. In the first phase, we run the approximation algorithm of Hosoda *et al.* and then that of Angluin *et al.* taking into account the edges already constructed by our algorithm. In the second phase, starting from scratch, we can run the two approximation methods in the reverse order analogously. Finally, we output the smallest of the union outputs produced in the two phases. See Fig. 2. Although this refinement may not always lead to a substantially better approximation compared to the hybrid method (see Theorem 6), it may be useful in practice.

Require: a vertex set V, edge construction costs $c(e)$, a set E' of edges e with $c(e) = 0$, and a set S connectivity constraints $S_1, ..., S_r$.
Ensure: A set of edges yielding an approximate solution to the Network Construction problem for the input instance.

```
S_{≤B} ← the set of all constraints in S with at most B vertices
S_{>B} ← the set of all constraints in S with more than B vertices
Use the method of Hosoda et al.from Fact 2 to solve the Network Construction problem with the set E' of edges having zero construction cost and the constraint set S_{≤B} by an edge set E'_1 ⊆ V^2
Set the construction costs of the edges in E'_1 to zero and use the method of Angluin et al. from Fact 1 to solve the Network Construction problem with the set E' ∪ E'_1 of edges having zero cost, and the constraint set S_{>B}, by an edge set E''_1 ⊆ V^2
Use the method of Angluin et al. from Fact 1 to solve the Network Construction problem with the set E' of edges having zero cost, and the constraint set S_{>B}, by an edge set E'_2 ⊆ V^2
Set the construction costs of the edges in E'_2 to zero and use the method of Hosoda et al. from 2 with the set E' ∪ E'_2 of edges having zero construction cost, and the constraint set S_{≤B}, by an edge set E''_2 ⊆ V^2
return the set in {E''_1, E''_2} of minimum total cost.
```

Fig. 2. The refined hybrid approximation algorithm for the Network Construction problem.

4.2. *Bounded diameter requirements*

One can naturally strengthen the connectivity requirements in the Network Construction or Extension problems by demanding that each constraint should induce a subgraph of the constructed network of diameter at most d, where $d \in [n-1]$ is given a priori (cf. [6]).

For instance, Chockler *et al.* studied the Network Construction problem in [6] using a different terminology. They considered the problem of constructing an optimal overlay (network) that for each topic (constraint) includes a dissemination tree composed of nodes interested in the topic (i.e., belonging to the constraint). One of the measures of the quality of such an overlay suggested on p. 116 of [6] is the diameter. Intuitively, having a low diameter is good because it means that two users interested in the same topic do not need to rely on many intermediate parties, which leads to more efficient communication and better performance.

We shall term the strengthened version of the Network Construction problem as the d-diameter Network Construction problem. In fact, the latter problem restricted to instances with a single constraint is already hard. The restriction can be simply rephrased as follows: given a vertex set V, edge costs $c(e)$ for potential edges in V^2, find a cheapest graph spanning V with diameter not exceeding d.

The d-diameter Network Construction problem restricted to single-constraint instances is known to be NP-hard already for $d = 2$ [14] (given an instance (V, E) of the minimum-cardinality-bounded-diameter edge addition problem with $D = 2$, studied and shown to be NP-hard in [14], set the cost of edges in E to zero and the cost of all other potential edges to 1 to obtain an equivalent instance of the 2-diameter Network Construction problem with single constraints). In contrast, when restricted to instances with uniform edge costs, this problem variant becomes trivial as any spanning star graph provides an optimal solution. In [5], Bilò *et al.* improved the approximation ratio for the related minimum-cardinality-bounded-diameter problem to $O(\log n)$, closing the asymptotic gap between lower and upper bounds on the approximability of this problem.

Analogously to the preceding sections, we can consider the d-diameter Network Construction problem with constraints of cardinality not exceeding B. By using an *auxiliary problem*, we can obtain a $(\binom{B}{2} - B + 2)\binom{B}{2}$ approximation in polynomial time for the B-constraint-bounded d-diameter Network Construction problem. The auxiliary problem is as follows.

For an instance of the B-constraint-bounded d-diameter Network Construction problem with a vertex set V, edge construction costs $c(e)$, a set E' of edges e with $c(e) = 0$, and connectivity constraints $S_1, \ldots, S_r$ find a minimum cost edge set $E'' \subseteq V^2 \setminus E'$ such that for $i = 1, \ldots, r$, if the diameter of the subgraph of $G' = (V, E)$ induced by S_i is larger than d then $E'' \cap S_i^2 \neq \emptyset$.

The following lemma provides an approximation algorithm for the auxiliary problem.

Lemma 7. *The auxiliary problem can be approximated within* $\binom{B}{2}$ *in polynomial time.*

Proof. Consider an instance of the auxiliary problem with a vertex set V, edge construction costs $c(e)$, a set E' of edges with zero construction cost, and connectivity constraints $S_1, \ldots, S_r$. We may assume w.l.o.g. that for $i = 1, \ldots, r$, the diameter of the subgraph of the graph $G' = (V, E')$ induced by S_i is larger than d since otherwise the constraint S_i can be disregarded. To solve the auxiliary problem, for $i = 1, \ldots, r$, form the set E_i of all edges in $S_i^2 \setminus E'$. The auxiliary problem is equivalent to finding a minimum weight subset of the set of all potential edges that hits all the sets $E_1, \ldots, E_r$, where the weights of the edges are equal to their construction costs. By our assumptions, for $i = 1, \ldots, r$, $|E_i| \leq \binom{B}{2}$ hold. Now it is sufficient to apply Fact 3 in order to obtain a $\binom{B}{2}$ approximation for the auxiliary problem in time linear in the total size of the family $\{E_1, \ldots, E_r\}$ and $|V|^2$. The latter size is in turn polynomial in the size of the input instance of the auxiliary problem. $\square$

Now, in order to provide an approximate solution to an instance of the B-constraint-bounded d-diameter Network Construction problem, we iterate the method of Lemma 7 as shown in Fig. 3.

Require: a vertex set V, edge construction costs $c(e)$, a set E' of edges e with $c(e) = 0$, and d-diameter constraints $S_1, ..., S_r$ of cardinality $\leq B$.
Ensure: A set of edges yielding an $(\binom{B}{2} - B + 2)\binom{B}{2}$ approximation to the d-diameter Network Construction problem for the input instance.

$E'' \leftarrow E'$
while there is a not yet satisfied constraint S_i **do**
Use the method of Lemma 7 to find an approximate solution $E''' \subseteq V^2 \setminus E''$ to the auxiliary problem.
$E'' \leftarrow E'' \cup E'''$ (E'' can be interpreted as the set of already constructed edges)
Remove all connectivity constraints S_i such that the subgraph of $G' = (V, E'')$ induced by S_i has diameter $\leq d$.
Set the construction costs of the edges in E''' to zero.
end while
return $E'' \setminus E'$

Fig. 3. The $(\binom{B}{2} - B + 2)\binom{B}{2}$ approximation algorithm for the B-constraint-bounded d-diameter Network Construction problem.

Theorem 8. *The B-constraint-bounded d-diameter Network Construction problem can be approximated within $(\binom{B}{2} - B + 2)\binom{B}{2}$ in polynomial time.*

Proof. We shall analyze the iterative method based on Lemma 7. Since for $i = 1, \ldots, r$, $|S_i| \leq B$, the subgraph of the original graph $G' = (V, E')$ induced by S_i can be completed by at most $\binom{B}{2}$ edges. Hence, at most $\binom{B}{2}$ iterations of the while block are sufficient. In fact, already $\binom{B}{2} - B + 2$ iterations are sufficient since in a graph with B vertices and at least $\binom{B}{2} - (B - 2)$ edges each pair of non-adjacent vertices has a common neighbor. Note that the cost of an optimal solution to any of the at most $\binom{B}{2} - B + 2$ auxiliary problems approximately solved in consecutive iterations of the while block cannot be greater than that of an optimal solution to the original B-constraint-bounded d-diameter Network Construction problem. Hence, the upper bound $(\binom{B}{2} - B + 2)\binom{B}{2}$ on the approximation factor of the iterative method follows from Lemma 7. $\square$

Hosoda *et al.* showed the B-constraint-bounded Network Construction problem to be APX-complete in [10]. Similarly, we can show the B-constraint-bounded d-diameter Network Construction problem to be APX-complete by using Theorem 8.

Theorem 9. *The B-constraint-bounded d-diameter Network Construction problem is APX-complete, when $B \geq 3$ and $d \geq 2$. It is also NP-hard to approximate beyond $(B - 2)$ factor and UGC-hard to approximate beyond $B - 1$, for $B \geq 3$ and $d \geq 2$.*

Proof. By Theorem 8, it is sufficient to show that the B-constraint-bounded d-diameter Network Construction problem is APX-hard in order to show the APX-completeness of this problem.

Consider an instance I of the Minimum Weight Hitting Set problem with a finite set U of elements $u_1, \ldots, u_r$ having uniform weights 1, and a family C of subsets of U, each of cardinality not exceeding $B-1$. Form a set V of vertices a and $v_1, v_1, \ldots, v_r$, where for $i = 1, \ldots, r$, the pair of vertices a, v_i corresponds to the element u_i. Set the construction cost of edges connecting two vertices v_i, v_j to zero, the construction cost of edges $\{a, v_i\}$ for $i = 1, \ldots, r$, to 1. Now, for each subset C_j in C, form the constraint $S_j = \{a\} \cup \{v_i | u_i \in C_j\}$. By our assumptions and the construction, $|S_j| \le B - 1 + 1 \le B$ holds. Note that if all edges of zero cost are constructed and for each constraint S_j at least one edge of the form $\{a, v_i\}$ (having cost 1), where $v_i \in C_j$, is constructed then constraints induce subgraphs of diameter at most d (in fact of diameter at most 2). Otherwise, at least one of the subgraphs induced by a constraint is disconnected. The construction of an edge $\{a, v_i\}$ corresponds to the insertion of the element u_i into the hitting set. Thus, the d-diameter Network Construction problem for the aforementioned instance becomes equivalent to the Minimum Weight Hitting Set problem for the instance I, in this setting. Hence, by Fact 4, we obtain APX-hardness of the B-constraint-bounded Network Construction problem. The APX-completeness follows now from Theorem 8. Finally, by Fact 5, we conclude that B-constraint-bounded d-diameter Network Construction problem is NP-hard to approximate beyond $(B-2)$ factor and UGC-hard to approximate beyond $B-1$ factor, for $B \ge 3$, $d \ge 2$. □

In the general case with unbounded constraints, straightforward greedy approaches do not seem to work. However, if the edge costs are uniform, we can obtain a large but still a nontrivial approximation factor in polynomial time by splitting the constraints into small and large ones, and using Theorem 8 to obtain an approximation for the former.

For an instance of the uniform cost d-diameter Network Construction problem with n vertices, consider the overlap graph whose vertices are in one-to-one correspondence with the input constraints and a pair of vertices is adjacent if and only if the corresponding constraints overlap. Note that if the overlap graph is connected then any solution to the d-diameter problem has to include at least $n-1$ edges. Otherwise, one can consider subproblems of the d-diameter problem constrained to the unions of constraints belonging to the same connected component of the overlap graph instead.

Theorem 10. *The uniform cost d-diameter Network Construction problem with n vertices and r constraints, $d \ge 2$, admits an $O((n \ln r)^{0.8})$ approximation, when a solution graph has to be connected.*

Proof. Split the constraints into small ones of size $\le (n \ln r)^{0.2}$ and the remaining large ones. We can apply the $(\binom{B}{2} - B + 2)\binom{B}{2}$ approximation to the small constraints obtaining an $(n \ln r)^{0.8}$ approximation. To satisfy the large constraints on more than $(n \ln r)^{0.2}$ vertices, we proceed as follows.

Let Q be the set of vertices involved in the large constraints. and let q stand for the cardinality of Q. Thus, we have a sequence of sets (i.e., constraints) $S_1, \ldots, S_k$, $k \leq r$, each of size at least $t = (n \ln r)^{0.2}$. We would like to hit all these sets using the elements from Q and for each vertex v in the hitting set to construct all edges connecting v with other vertices in Q. To form our hitting set, we use the standard greedy algorithm. Since the set size is at least t, each time at least t/n fraction of the sets will be hit.

This is easily seen when the sets are disjoint, i.e., when each vertex occurs in at most one of the sets. Suppose that the maximum number of the sets that a vertex in Q can occur is $l \geq 1$. Then, the union of the sets forms a multiset of size at most nl and consequently their number k does not exceed $\frac{nl}{t}$. Hence, the greedy heuristic picks an element hitting at least t/n fraction of the sets.

As a result, the greedy algorithm produces a hitting set of size $\frac{n}{t} \ln r = (n \ln r)^{0.8}$. Thus, the total cost of the constructed edges to satisfy the large constraints is at most $(n \ln r)^{0.8}(q-1)$. Since the cost of an optimal solution is at least $q-1$ by the theorem assumption, we obtain an $(n \ln r)^{0.8}$ approximation for the large constraints. □

5. Final Remarks

It would be useful to tighten the upper and lower competitiveness bounds on the online version of the B-constraint-bounded Network Construction problem. It would be especially interesting to know if the factor that is logarithmic in n can be removed from the upper bounds.

As mentioned in Sec. 4.2, straightforward greedy approaches do not seem to work for the d-diameter Network Construction problem with unbounded constraints. One reason for this is that natural candidates for potential functions in greedy methods (e.g., the number of pairs of vertices within distance d of each other) seem to lack the submodularity property. It is an interesting question if it is possible to achieve a reasonable approximation factor for this problem in the general case, at least when edge costs are uniform.

Finally, it would be interesting to consider fixed-parameter algorithms for the Network Construction Problem parameterized by B, r, and d.

Acknowledgments

We would like to thank Tatsuya Akutsu and Natsu Nakajima for introducing us to the problem studied in this paper, and Oleg Verbitsky for inspiring questions. We are also very grateful to anonymous reviewers for their valuable comments and suggestions. The research was supported in part by Swedish Research Council grant 621-2017-03750 and KAKENHI grant 22H03550.

References

[1] N. Alon, B. Awerbuch, Y. Azar, N. Buchbinder and J. Naor, The online set cover problem, *SIAM J. Comput.* **39**(2) (2009) 361–370.

[2] D. Angluin, J. Aspnes and L. Reyzin, Network construction with subgraph connectivity constraints, *J. Combinatorial Optimization* **29** (2015) 418–432.

[3] G. Ausiello, A. D'Atri and M. Protasi, Structure preserving reductions among convex optimization problems, *J. Comput. Syst. Sci.* **21** (1980) 136–153.

[4] R. Bar-Yehuda and S. Even, A linear-time approximation algorithm for the weighted vertex cover problem, *J. Algorithms* **2**(2) (1981) 198–203.

[5] D. Bil, L. Gual and G. Proietti, Improved approximability and non-approximability results for graph diameter decreasing problems, *Theoretical Comput. Sci.* **417** (2012) 12–22.

[6] G. V. Chockler, R. Melamed, Y. Tock and R. Vitenberg, Constructing scalable overlays for pub-sub with many topics, In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Principles of Distributed Computing (PODC 2007)*, pp. 109–118, 2007.

[7] I. Dinur, V. Guruswami, S. Khot and O. Regev, A new multilayered pcp and the hardness of hypergraph vertex cover, *SIAM J. Comput.* **34**(5) (2005) 1129–1146.

[8] M. R. Garey and D. S. Johnson, Computers and intractability — A guide to the theory of NP-completeness, *W. H. Freeman and Company* New York, 1979.

[9] D. S. Hochbaum, Approximation algorithms for the set covering and vertex cover problems, *SIAM J. Comput.* **11**(3) (1982) 555–556.

[10] J. Hosoda, J. Hromkovič, T. Izumi, H. Ono, M. Steinová and K. Wada, On the approximability and hardness of minimum topic connected overlay and its special instances, *Theoretical Computer Science* **429** (2012) 144–154.

[11] S. Khot, On the power of unique 2-prover 1-round games, *Proc. of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, pp. 767–775, 2002.

[12] S. Khot and O. Regev, Vertex cover might be hard to approximate to within 2-epsilon, *J. Comput. Syst. Sci.* **74**(3) (2008) 335–349.

[13] E. Korach and M. Stern, The clustering matroid and the optimal clustering tree, *Mathematical Programming* **98** (2003) 385–414.

[14] C. Li, S. T. McCormick and D. Simchi-Levi, On the minimum-cardinality-bounded-diameter and the bounded-cardinality-minimum-diameter edge addition problems, *Oper. Res. Lett.* **11** (1992) 303–308.

[15] N. Nakajima, M. Hayashida, J. Jansson, O. Maruyama and T. Akutsu, Determining the minimum number of protein-protein interactions required to support known protein complexes, *PLOS ONE* **13**(4), Article e0195545, 2018.

[16] C. H. Papadimitriou and M. Yannakakis, Optimization, approximation and complexity classes, *J. Computer and System Sciences* **43** (1991) 425–440.