

Efficient Distributed Workload (Re-)Embedding

MONIKA HENZINGER*, STEFAN NEUMANN*[†], and STEFAN SCHMID*, Faculty of Computer Science, University of Vienna, Austria

Modern networked systems are increasingly reconfigurable, enabling *demand-aware* infrastructures whose resources can be adjusted according to the workload they currently serve. Such dynamic adjustments can be exploited to improve network utilization and hence performance, by *moving* frequently interacting communication partners closer, e.g., collocating them in the same server or datacenter. However, dynamically changing the embedding of workloads is algorithmically challenging: communication patterns are often not known ahead of time, but must be *learned*. During the learning process, overheads related to unnecessary moves (i.e., re-embeddings) should be minimized. This paper studies a fundamental model which captures the tradeoff between the benefits and costs of dynamically collocating communication partners on ℓ servers, in an online manner. Our main contribution is a distributed online algorithm which is asymptotically almost optimal, i.e., almost matches the lower bound (also derived in this paper) on the competitive ratio of any (distributed or centralized) online algorithm. As an application, we show that our algorithm can be used to solve a distributed union find problem in which the sets are stored across multiple servers.

1 INTRODUCTION

Along with the trend towards more *data centric* applications (e.g., online services like web search, social networking, financial services as well as emerging applications such as distributed machine learning [1, 2]), comes a need to *scale out* such applications, and distribute the workload across multiple servers or even datacenters. However, while such parallel processing can improve performance, it can entail a non-trivial load on the interconnecting network. Indeed, distributed cloud applications, such as batch processing, streaming, or scale-out databases, can generate a significant amount of network traffic [3].

At the same time, emerging networked systems are becoming increasingly flexible and thereby provide novel opportunities to mitigate the overhead that distributed applications impose on the network. In particular, the more flexible and dynamic resource allocation (enabled, e.g., by virtualization) introduces a vision of *workload-aware* infrastructures which optimize themselves to the demand [4]. In such infrastructures, communication partners which interact intensively, may be *moved* closer (e.g., colocated on the same server, rack, or datacenter) in an adaptive manner, depending on the demand. This “re-embedding” of the workload allows to keep communication local and reduce costs. Indeed, empirical studies have shown that communication patterns in distributed applications feature much locality, which highlights the potential of such self-adjusting networked systems [5–7].

* Authors are ordered alphabetically.

[†] Contact author.

Authors' address: Monika Henzinger; Stefan Neumann, stefan.neumann@univie.ac.at; Stefan Schmid, Faculty of Computer Science, University of Vienna, Währinger Strasse 29, Vienna, 1090, Austria.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

XXXX-XXXX/2019/1-ART

<https://doi.org/0000001.0000001>

However, leveraging such resource reconfiguration flexibilities to optimize performance, poses an algorithmic challenge. *First*, while collocating communication partners reduces communication cost, it also introduces a *reconfiguration cost* (e.g., due to virtual machine migration). Thus, an algorithm needs to strike a balance between the benefits and the cost of such reconfigurations. *Second*, as workloads and communication patterns are usually not known ahead of time, reconfiguration decisions need to be made in an *online* manner, i.e., without knowing the future. We are hence in the realm of online algorithms and competitive analysis.

This paper studies the fundamental tradeoff underlying the optimization of such workload-aware reconfigurable systems. In particular, we consider the design of an online algorithm which, without prior knowledge of the workload, aims to minimize communication cost by performing a small number of *moves* (i.e., migrations). In a nutshell (more details will follow below), we consider a *communication graph* between n vertices (e.g., virtual machines) which can be perfectly partitioned among a set of ℓ servers (resp. racks or datacenters) of a given *capacity*. We assume that the *communication patterns*, which partition the communication graph, consist of n/ℓ vertices and that once the whole communication graph was revealed, each server must contain exactly one communication pattern.

The communication graph is initially unknown and revealed to the algorithm in an online manner, edge-by-edge, by an adversary who aims to maximize the cost of the given algorithm. The cost here consists of *communication cost* and *moving cost*: The algorithm incurs one unit cost if the two endpoints (i.e., communication partners) of the request belong to different servers. After each request, the algorithm can reconfigure the infrastructure and move communication endpoints from one server to another, essentially *repartitioning* the communication partners; however, each move incurs a cost of $\alpha > 1$.

In other words, this paper considers the problem of *learning a partition*, i.e., an optimal assignment of communication partners to servers, at low communication and moving cost. Interestingly, while the problem is natural and fundamental, not much is known today about the algorithmic challenges underlying this problem, except for the negative result that no good competitive algorithm can exist if communication partners can change arbitrarily over time [8]. This lower bound motivates us, in this paper, to focus on the online *learning variant* where the communication partners are unknown but fixed. At the same time, as we will show, the problem features interesting connections to several classic problems. Specifically, the problem can be seen as a *distributed* version of classic online caching problems [9] or an *online* version of the k -way partitioning problem [10].

1.1 Our Contributions

We initiate the study of a fundamental problem, how to learn and re-embed workload in an online manner, with few moves. We make the following main contributions.

We present a distributed $O((\ell \log \ell \log n)/\varepsilon)$ -competitive online algorithm for servers of capacity $(1 + \varepsilon)n/\ell$, where $\varepsilon \in (0, 1/2)$. We allow the servers to have $\varepsilon n/\ell$ more space than is strictly needed to embed its corresponding communication pattern (which is of size n/ℓ); we denote this additional space as *augmentation*. Such augmentation is also needed, as our lower bounds discussed next show.

We show that there are inherent limitations of what online algorithms can achieve in our model: We derive a lower bound of $\Omega(1/\varepsilon + \log n)$ on the competitive ratio of any deterministic online algorithm given servers of capacity at least $(1 + \varepsilon)n/\ell$. This lower bound has several consequences: (1) To obtain $O(\log n)$ -competitive algorithms, the servers must have $\Omega(n/(\ell \log n))$ augmentation. (2) If the servers have $\Omega(n/\ell)$ augmentation (e.g., each server has 10% more capacity than the size of its communication pattern), our algorithm is optimal up to an $O(\ell \log \ell)$ factor. Thus, our results are particularly interesting for large servers, e.g., in a wide-area networking context where there is

usually only a small number of datacenters where communication partners can be collocated (e.g., $\ell = 20$): if each datacenter (“server”) has augmentation $0.1 \cdot n/\ell$, our algorithm is optimal up to constant factors.

The distributed algorithms we present not only provide good competitive ratios but they are also highly efficient w.r.t. the network traffic they cause. In fact, we show that for $\ell = O(\sqrt{\varepsilon n})$ servers, running the algorithms introduces only little overhead in network traffic and that this overhead is asymptotically negligible (see Section 5.1).

While the previous algorithms require exponential time, we also present polynomial time algorithms at the cost of a slightly worse competitive ratio of $O((\ell^2 \log n \log \ell)/\varepsilon^2)$ in Section 5.2.

As a sample application of our newly introduced model we present a distributed union find data structure [11, 12] (also known as disjoint-set data structure or merge-find data structure) in Section 7.1: There are n items from a universe which are distributed over ℓ servers; each server can store at most $(1 + \varepsilon)n/\ell$ items and each item belongs to a unique set. The operation *union* allows to merge two sets. In our setting, we require that items from the same set must be assigned to the same server. To reduce the network traffic, our goal is to minimize the number of item moves during union operations. For example, when two sets are merged which are assigned to different servers, then the items of one of the sets must be reassigned to another server. We compare against an optimal offline algorithm which knows the initial assignment of all items and all union operations in advance. We obtain the same competitive ratios as above. We believe that this distributed union find data structure will be useful as a subroutine for several problems such as merging duplicate websites in search engines [13].

We also show that our algorithms solve an online version of the k -way partition problem in Section 7.2.

1.2 Organization

We introduce our model formally in Section 2. To ease the readability, we first explore centralized online algorithms that efficiently collocate communication patterns for $\ell = 2$ servers in Section 3, and then study the general case of $\ell > 2$ servers in Section 4. In Section 5 we show how the previously derived centralized algorithms can be made distributed and how the algorithm can be implemented in polynomial time at the cost of a slightly worse competitive ratio. We provide the lower bounds in Section 6. Section 7 provides a distributed union find data structure and a result for online k -way partitioning; these problems serve as sample applications of the problem we study. After reviewing related work in Section 8, we conclude our contribution in Section 9.

2 MODEL

We start by formally introducing the model which we will be studying in this paper. We consider a set of vertices V (e.g., a set of virtual machines) which interact according to an initially unknown communication pattern, which can be represented as a communication graph $G = (V, E)$ with $n = |V|$ vertices and $m = |E|$ edges. The vertices of G are partitioned into ℓ sets $V_0, \dots, V_{\ell-1}$ where each V_i , forming a connected communication component (the workload), has size¹ n/ℓ ; the connected components of G coincide with the sets V_i . The sets V_i are the communication patterns which need to be recovered by the online algorithm, henceforth called *ground truth components*.

The communicating vertices V need to be assigned to ℓ servers $S_0, \dots, S_{\ell-1}$. Accordingly, we define an *assignment* (the embedding) which is a function from the vertices to the servers. The *load* of a server S_j is the number of vertices that are assigned to it. An assignment is *valid* if each server

¹Note that in general n/ℓ is not always an integer and we would have to take rounding into account. However, we ignore this technicality for better readability of the paper.

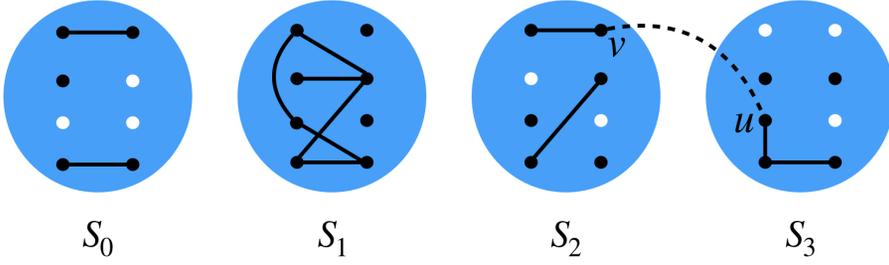


Fig. 1. An illustration of the model we consider. In the picture there are $\ell = 4$ servers each depicted by a blue circle. Vertices assigned to a server are represented by black dots whereas white dots represent unused server capacities. Note that there are $n = 24$ vertices and each server has capacity $(1 + \varepsilon)n/\ell = 8$ for $\varepsilon = 1/3$. In the picture, server S_0 has load 5 and server S_1 has load 8. When two vertices communicated, we draw an edge between them depicted by a black line. Observe how this naturally gives rise to connected components and note that S_1 contains a ground truth component of size $n/\ell = 6$. If the adversary were to insert the edge (u, v) next, the algorithm could, for example, move the connected component containing v to S_3 at cost 2α .

has load at most $n/\ell + K$ and we call $n/\ell + K$ the *capacity* of the servers and K the *augmentation*. If $K = 0$, the total server capacity exactly matches the number of vertices. The *available capacity* of a server is the difference between the server's capacity and its load. An assignment is *perfectly balanced* if each server has load exactly n/ℓ . We assume that when the algorithm starts, we have a perfectly balanced assignment. We will write $V(S_j)$ to denote the set of vertices assigned to server S_j and $V_{\text{init}}(S_j)$ for the set of vertices *initially* assigned to server S_j . We say that an assignment is a *perfect partitioning* if it satisfies $\{V(S_0), \dots, V(S_{\ell-1})\} = \{V_0, \dots, V_{\ell-1}\}$, i.e., the vertices on the servers coincide with the connected components of G .

The communication graph $G = (V, E)$ is revealed by an adversary in an *online manner*, as a sequence of edges $\sigma = (e_1, \dots, e_r)$, where r denotes the number of communication requests and $e_i \in E$ for each i . Note that the adversary can only provide edges which are present in E and that each edge can appear multiple times in the sequence of edges. We assume that the sequence of the edges provided by the adversary reveals the ground truth components V_i , i.e., after having seen all edges in σ the algorithm can compute the connected components of G which (by assumption) coincide with the ground truth components V_i . We present an illustration of the model in Figure 1.

Now an online algorithm must iteratively change the assignment such that eventually the assignment is a perfect partitioning.

The reassignment needs to be done while minimizing certain communication and migration cost. If an edge $e = (u, v)$ provided by the adversary has both endpoints in the same server S_i at the time of the request, an algorithm incurs no costs. If u and v are in different servers S_i and S_j , then their communication cost is 1. Reassigning, i.e., *moving*, a vertex u from a server S_i to a server S_j costs $\alpha > 1$.

When measuring the cost of an online algorithm, we will compare against an optimal offline algorithm denoted by OPT. OPT has *a priori* knowledge of the communication graph $G = (V, E)$ as well as the given the sequence of all edges $\sigma = (e_1, \dots, e_r)$. In other words, OPT can compute the assignment of vertices to servers which provides the minimum migration cost from the initial assignment.

Now let the cost paid by an online algorithm be denoted by ON and let the cost of the optimal offline algorithm be denoted by OPT. We consider the design of an online algorithm ON which minimizes the (strict) *competitive ratio* defined as $\frac{\text{ON}}{\text{OPT}}$.

The Role of Connected Components. We will briefly discuss how connected components are induced by subsequence of σ and how we will treat connected components in our algorithms. We then give a reduction which helps us to avoid considering communication costs in our proofs.

Recall that the adversary provides a sequence of edges σ to an algorithm in an online manner. As this happens, an algorithm can keep track of all edges it has seen so far. Let this set of edges be E' . Using the edges in E' , the algorithm can compute the connected components C_1, \dots, C_q which are induced by E' . Here, q denotes the current number of connected components.

To obtain a better understanding of the relationship between the connected components C_i and the ground truth components V_j , we make four observations: (1) When the algorithm starts, all connected components $C_i = \{v_i\}$ only consist of single vertices (because σ has not yet revealed any edges). (2) When a previously unknown edge $e = (u, v)$ is revealed which has its endpoints in different connected components C_u and C_v , these connected components get merged. (3) Suppose a subsequence of σ induces $q > \ell$ connected components C_i (i.e., σ has not yet revealed the whole graph G). Then for each ground truth component V_j there exists a subset $\mathbb{C} \subset \{C_1, \dots, C_q\}$ of the connected components such that $V_j = \bigcup_{C \in \mathbb{C}} C$. (4) When an algorithm terminates (and, hence, σ revealed all edges in E), there exists a one-to-one correspondence between the connected components C_i and the ground truth components V_j .

By assumption on the input from the adversary, when all of σ was revealed, E' reveals the ground truth components $V_0, \dots, V_{\ell-1}$. Thus, in total there will be exactly $n - \ell$ edges connecting vertices from different connected components.

All of the algorithms we consider in this paper have the property that they always assign vertices of the same connected component to the same server. This property implies that the communication cost paid by such an algorithm is bounded by its moving cost (we prove this in the following lemma). Hence, in the rest of the paper we only need to bound the moving costs of our algorithms to obtain a bound on their total costs.

LEMMA 1. *Suppose an algorithm \mathcal{A} always assigns all vertices of the same connected component to the same server and pays \mathcal{C} for moving vertices. Then its communication cost is at most \mathcal{C} . Furthermore, its total cost is at most $2\mathcal{C}$.*

PROOF. Suppose the adversary provides an edge (u, v) . We consider two cases. *Case 1:* u and v are assigned to the same server. Then \mathcal{A} does not pay any communication costs. *Case 2:* u and v are assigned to connected components C_u and C_v on different servers. Then the algorithm needs to pay 1 communication cost. However, in this case \mathcal{A} must move C_u or C_v to a different server at the cost of at least $\alpha > 1$. Hence, the moving cost is larger than the communication cost. We conclude that \mathcal{A} 's total communication cost is at most \mathcal{C} . By summing the two quantities, we obtain the second claim of the lemma. \square

While in Lemma 1 we have shown that algorithms which always collocate connected components immediately are efficient w.r.t. their total cost, in Section 6.1 we show that any efficient algorithm must satisfy a similar (slightly more general) property.

Throughout the rest of the paper, we write $|C|$ to denote the number of vertices in a connected component C . For a vertex u , we write C_u to denote the connected component C which contains u .

3 ONLINE PARTITION FOR TWO SERVERS

In this section, we consider the problem of learning a communication graph with few moves with *two* servers. As we will see later, the concepts introduced in this section will be useful when solving the problem with $\ell > 2$ servers. We derive the following result.

THEOREM 2. *Consider the setting with two servers of capacity $(1 + \epsilon)n/2$ for $\epsilon \in (0, 1)$, i.e., the augmentation is $\epsilon n/2$. Then there exists an algorithm with competitive ratio $O((\log n)/\epsilon)$.*

The proof is organized as follows. We first characterize the optimal solution by OPT in Section 3.1. We then present an algorithm which is efficient whenever OPT incurs “significant cost”, in Section 3.2. In Section 3.3, we describe an algorithm which is efficient whenever the solution by OPT is “cheap”. We prove Theorem 2 via a combination of the two algorithms in Section 3.4.

3.1 Costs of OPT

The following lemma gives a precise characterization of the cost paid by OPT in the two server case. It introduces a parameter Δ which equals the number of vertices moved by OPT and which we will be using throughout the rest of this section.

LEMMA 3. *Suppose $\ell = 2$ and the vertices initially assigned to the servers S_i are given by the sets $V_{\text{init}}(S_i)$ for $i = 0, 1$. Then the cost of OPT is $2\alpha\Delta$, where*

$$\Delta = \min\{|V_{\text{init}}(S_0) \cap V_0|, |V_{\text{init}}(S_0) \cap V_1|\}.$$

It follows immediately that $\Delta \leq n/4$ (as $|V_{\text{init}}(S_0)| = n/2$).

PROOF. Recall that our model forces OPT to provide a final assignment satisfying $\{V(S_0), V(S_1)\} = \{V_0, V_1\}$, i.e., OPT must produce a final assignment which coincides with the ground truth components (even if paying for each communication request individually and not relocating any vertices might be cheaper). Thus, we can assume that OPT performs all vertex moves in the beginning, to avoid paying any communication cost. Since the edge sequence $\sigma = (e_1, \dots, e_r)$ provided by the adversary is assumed to reveal the connected components V_0 and V_1 , OPT can compute V_0 and V_1 before it performs any moves.

As there are only two servers, one of them must contain at least half of the vertices from V_0 in the initial assignment. Now let us first assume that this server is S_0 ; this setting is illustrated in Figure 2. In this case, OPT can move the Δ vertices in $V_{\text{init}}(S_0) \cap V_1$ to S_1 and those in $V_{\text{init}}(S_1) \cap V_0$ to S_0 . It is easy to verify that this yields an assignment satisfying $\{V(S_0), V(S_1)\} = \{V_0, V_1\}$ and that the moving cost is minimized. Further, the cost for this reassignment is exactly $2\alpha\Delta$.

The second case where S_1 contains more than half of the vertices from V_0 in the initial assignment is symmetric. \square

While in Lemma 3 we have presented the lower bound w.r.t. server S_0 , we could also express the lower bound in terms of server S_1 . We then obtain the following equality:

$$\Delta = \max_{i=0,1} \min_{j=0,1} |V_{\text{init}}(S_i) \cap V_j|.$$

3.2 The Small–Large–Rebalance Algorithm

A natural idea to obtain a small number of vertex moves is to proceed as follows. Whenever two vertices u and v belonging to different connected components communicate, the algorithm merges their connected components. If the two components were already assigned to the same server, no vertex moves are required. If u and v are assigned to different servers, we move the smaller connected component to the server of the larger connected component. This algorithm is efficient in that it never performs more than $O(n \log n)$ vertex moves (see Lemma 4).

However, the algorithm could require much augmentation, as it does not account for server capacities. Thus, we propose the following extension called the *Small–Large–Rebalance Algorithm*: Whenever a server exceeds its capacity, the algorithm computes a perfectly balanced assignment

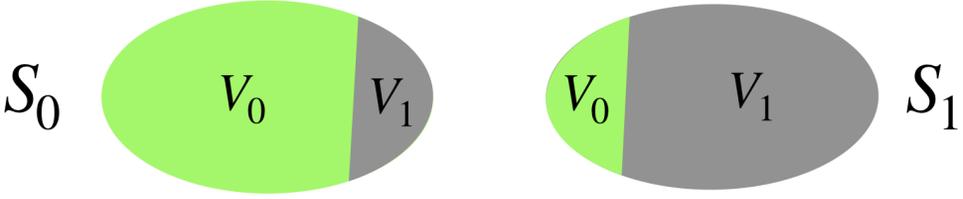


Fig. 2. The initial assignment considered in the proof of Lemma 3. The green and grey areas of the servers correspond to subsets of V_0 and V_1 . Server S_0 (S_1) contains most of the vertices from V_0 (V_1). Here, OPT would move the green part from S_1 to S_0 and the grey part from S_0 to S_1 .

of the vertices which respects the previously observed connected components; we call this a *rebalancing step*. We provide pseudocode in Algorithm 1.

Section 5.2.1 shows how such a rebalancing step can be implemented in $O(n^2)$ time. Later, we show that there can be at most $O((\log n)/\epsilon)$ such rebalancing steps which implies that the total running time Algorithm 1 is $O((n^2 \log n)/\epsilon)$.

Note that Algorithm 1 also works in the setting with ℓ servers for $\ell > 2$. We will analyze this more general algorithm in Section 4.4.

3.2.1 Analysis. To analyze Algorithm 1, we first consider the algorithm from the first paragraph which does not have the rebalancing step. When the algorithm moves a smaller component to the server of a larger component, we call this a *small-to-large step*.

LEMMA 4. *Consider the algorithm which always moves the smaller connected component to the server of the larger connected component when it obtains an edge between vertices from different connected components. The algorithm moves each vertex at most $O(\log n)$ times. Its total number of vertex moves is $O(n \log n)$.*

PROOF. Consider any vertex v . We use the following accounting: Whenever v is in the a smaller component that is moved, add a token to v . Now observe that whenever v gains a token, the size of its component at least doubles. This implies that v can be in the smaller component only $O(\log n)$ times. Thus, v cannot accumulate more than $O(\log n)$ tokens. Since this holds for each of the n vertices, the total number of moves is $O(n \log n)$. \square

The following lemma provides the analysis for Algorithm 1 which performs small-to-large steps and rebalancing steps.

LEMMA 5. *Suppose both servers have capacity $(1 + \epsilon)n/2$, i.e., the augmentation is $\epsilon n/2$ for $\epsilon \in (0, 1)$. Then Algorithm 1 performs $O((\log n)/\epsilon)$ rebalancing steps and $O((n \log n)/\epsilon)$ vertex moves.*

PROOF. We prove the bound on the number of vertex moves; the claim about the number of rebalancing steps is proved along the way. Note that all vertex moves performed by the algorithm originate from either small-to-large steps or from rebalancing steps. We bound the number of each of these vertex moves separately.

Note that the token-based argument from Lemma 4 still applies to the small-to-large steps of Algorithm 1. This implies that the total number of vertex moves due small-to-large steps is $O(n \log n)$.

Now consider the vertex moves caused by the rebalancing steps and recall that the initial assignment is perfectly balanced. Whenever a server exceeds its load, the small-to-large steps of

Algorithm 1 The Small–Large–Rebalance Algorithm

Input: A sequence of edges $\sigma = (e_1, \dots, e_r)$

```

1: procedure SMALLLARGEREBALANCE( $e_1, \dots, e_r$ )
2:   for  $i = 1, \dots, r$  do
3:      $(u, v) \leftarrow e_i$ 
4:     if  $C_u$  and  $C_v$  are not assigned to the same server then
5:        $\triangleright$  We must move  $C_u$  and  $C_v$  to the same server.
6:       Assume w.l.o.g. that  $|C_u| \leq |C_v|$ 
7:       if the server of  $C_v$  has available capacity  $|C_u|$  then
8:         Move  $C_u$  to the server of  $C_v$ 
9:        $\triangleright$  Small-to-large step
10:      else
11:         $\triangleright$  Rebalancing step
12:        Move to a perfectly balanced assignment respecting the connected components
13:      Merge  $C_u$  and  $C_v$ 

```

the algorithm must have moved at least $\varepsilon n/2$ vertices (because the augmentation of one of the servers is exceeded). This can only happen $O((n \log n)/(\varepsilon n)) = O((\log n)/\varepsilon)$ times since the total number of vertex moves due to small-to-large steps is $O(n \log n)$. Hence, the number of rebalancing steps is at most $O((\log n)/\varepsilon)$. Since each rebalancing step performs $O(n)$ vertex moves, the lemma follows. \square

3.2.2 More Efficient Rebalancing. We next propose a better rebalancing strategy which makes Algorithm 1 more efficient. So far, we used $\Theta(n)$ vertex moves for each rebalancing operation at the cost of $\Theta(\alpha n)$. We now bring the rebalancing cost down to $O(\text{OPT})$.

We adjust Algorithm 1 in the following way: Instead of rebalancing by taking *any* perfectly balanced assignment respecting the connected components (Line 11), we choose a perfectly balanced assignment respecting the connected components *which minimizes the number of vertex moves from the initial solution*. We call such an assignment *cheap*.

To find a cheap assignment, the algorithm could simply do the following: (1) Recall the initial assignment. (2) Exhaustively enumerate all perfectly balanced assignments respecting the connected components. (3) Among all of these assignments find one which is cheap. While such a simple algorithm can in principle be computationally costly, we can here exploit the online model of computation which allows us unlimited computational power. In Section 5.2 we show how less efficient rebalancing strategies can be implemented in polynomial time and we obtain slightly worse competitive ratios.

With the improved rebalancing strategy, we obtain Proposition 6.

PROPOSITION 6. *Suppose all servers have capacity $(1 + \varepsilon)n/2$, $\varepsilon > 0$. Then the number of vertex reassignments performed by Algorithm 1 with more efficient rebalancing is $O(n \log n + (\Delta \log n)/\varepsilon)$, where Δ is the number of vertex moves used by OPT.*

PROOF. First, note that the number of vertex moves for moving smaller components to larger components (Line 8) is $O(n \log n)$, by exactly the same arguments used in the proof of Lemma 5.

Second, we bound the number of vertex moves required for the rebalancing operations. Whenever the algorithm needs to rebalance, we can assume (for the sake of the analysis) that the algorithm makes the following three steps: (1) Roll back all changes done by small-to-large moves (Line 8) *since the last rebalancing operation*. Thus, after rolling back we have the same assignment as after the

last rebalancing operation. (2) Roll back to the initial assignment (by undoing the last rebalancing operation). (3) Move to a cheap assignment.

Observe that Step (1) and (2) of the previous three step procedure increase the number of vertex moves only by a constant factor compared to when the algorithm does not roll back: In total, Step (1) only adds additional $O(n \log n)$ vertex moves because each small-to-large move is undone exactly once. Step (2) only doubles the number of vertex moves for moving to cheap assignments as each rebalancing is only undone once.

Thus, we can complete the proof if we can show that the total number of vertex moves for moving from the initial assignment to the cheap assignments is bounded by $O((\Delta \log n)/\epsilon)$.

By Lemma 5, the number of rebalancing steps is bounded by $O((\log n)/\epsilon)$. Now we argue that for moving from the initial solution to each cheap assignment, the rebalancing moves at most $O(\Delta)$ vertices: Every time the algorithm computes a cheap rebalancing, the final solution obtained by OPT is a perfectly balanced assignment respecting the connected components. Thus, the number of vertex moves to obtain a cheap rebalancing is bounded by the number of moves performed by OPT which is $O(\Delta)$. This finishes the proof. \square

3.3 The Majority Voting Algorithm

We now present an algorithm which works well whenever the cost paid by OPT is small, i.e., when OPT only needs to move few vertices. The issue with Algorithm 1 from Section 3.2 is that during its execution, it might deviate much from the initial assignment (and thus move many vertices). The following algorithm has the property that it always stays close to the initial assignment.

For ease of readability, we will often refer to the two servers as the *left* and *right* servers, respectively, instead of calling them S_0 and S_1 .

Our algorithm starts by coloring vertices on the left server yellow and on the right server black. Throughout the execution of the algorithm, the vertices will keep this initially assigned color. The algorithm then follows the idea of always moving the smaller connected component to the server of the larger connected component; we will refer to this as *small-to-large step*. To stay close to the initial assignment, whenever the number of vertices in a newly merged connected component surpasses a power of 2, the algorithm performs a majority vote and moves the component to the server where more of its vertices originate from. More formally, we say that a set of vertices (e.g., a connected component) has a *yellow (black) majority* if it contains more yellow (black) vertices than black (yellow) vertices. In the majority voting step, the algorithm moves a component with a yellow (black) majority which is currently on the right (left) server to the left (right) server. The pseudocode for this procedure is stated in Algorithm 2.²

The reason for introducing the majority voting step is that it keeps the assignments produced by the algorithm during its runtime close to the initial assignment. Due to this property, we can show that the cost of Algorithm 2 is always close to the cost of OPT. The formal guarantees are stated in Proposition 7.

PROPOSITION 7. *Let Δ be the number of vertex moves performed by OPT (see Section 3.1). Then Algorithm 2 is $O(\log n)$ -competitive and the load of both servers is bounded by $n/2 + 4\Delta$.*

We devote rest of this subsection to the proof of the proposition. We start bounding the augmentation. For the proofs recall that V_0 and V_1 are the ground truth connected components of G .

²Note that in Algorithm 2 the following is possible when a component C_u is merged with a component C_v : C_u is moved from S to S_v due to a small-to-large step and immediately after that $C_u \cup C_v$ is moved back to S due to a majority-voting step. Thus, it would be more efficient to compute the result of the majority-voting step earlier and to move C_v to S immediately (without ever moving C_u to S_v). This modification would be slightly more efficient but it would affect the competitive ratio of the algorithm only by at most a constant factor. Thus, to simplify our analysis, we ignore this modification.

Algorithm 2 The Majority Voting Algorithm**Input:** A sequence of edges $\sigma = (e_1, \dots, e_r)$

```

1: procedure MAJORITYVOTING( $e_1, \dots, e_r$ )
2:   Color all vertices assigned to the left server yellow and all vertices assigned to the right
   server black
3:   for  $i = 1, \dots, r$  do
4:      $(u, v) \leftarrow e_i$ 
5:     Suppose w.l.o.g. that  $|C_u| \leq |C_v|$ 
6:     if  $C_u$  and  $C_v$  are on different servers then
7:       Move  $C_u$  to the server of  $C_v$  ▷ Small-to-large step
8:     Merge  $C_u$  and  $C_v$ 
9:     if there exists an  $i \in \mathbb{N}$  s.t.  $|C_u| < 2^i$ ,  $|C_v| < 2^i$  and  $|C_u \cup C_v| \geq 2^i$  then ▷ Majority
   voting step
10:      if  $C_u \cup C_v$  has a yellow majority then
11:        Move  $C_u \cup C_v$  to the left server
12:      if  $C_u \cup C_v$  has a black majority then
13:        Move  $C_u \cup C_v$  to the right server

```

In the following we are interested in what happened to a connected component since its last majority vote. To this end, we decompose it into a sequence of smaller connected components such that first a majority vote is performed and after that, only small-to-large steps are performed. For all of these small-to-large steps, the component will stay on the server that was picked by the majority vote. The following definition makes this notion formal.

DEFINITION 8 (DOUBLING DECOMPOSITION). *Let C be a connected component and let $s \in \mathbb{N}$ be such that $2^s \leq |C| < 2^{s+1}$. Consider k disjoint sets of vertices $C_i \subseteq V$ and let $\mathbb{C}_j = \bigcup_{i=1}^j C_i$ for $j = 1, \dots, k$. A sequence (C_1, \dots, C_k) is a doubling decomposition of C if the following properties hold:*

- (1) $C = \mathbb{C}_k = \bigcup_{i=1}^k C_i$,
- (2) during the execution of the algorithm, first $\mathbb{C}_1 \cup C_2$ are merged, then $\mathbb{C}_2 \cup C_3$ are merged, and, more generally, $\mathbb{C}_{i-1} \cup C_i$ is merged before $\mathbb{C}_i \cup C_{i+1}$,
- (3) for each $i = 1, \dots, k-1$, $|\mathbb{C}_i| \geq C_{i+1}$ and the algorithm moves C_{i+1} to the server of \mathbb{C}_i ,
- (4) $|C_1| < 2^s$ and $|\mathbb{C}_2| = |C_1 \cup C_2| \geq 2^s$.

Note that when considering a doubling decomposition, there will be exactly one majority-vote for the components \mathbb{C}_j – the one after C_1 and C_2 are merged. Thus, C and all \mathbb{C}_j , $j \geq 2$, will be assigned to the server that was picked in the majority vote of $C_1 \cup C_2$.

The following lemma shows that doubling decompositions are indeed well-defined. Its proof provides the construction of a doubling decomposition for a given connected component.

LEMMA 9. *Let C be a connected component. Then there exists a doubling decomposition (C_1, \dots, C_k) for C .*

PROOF. Suppose (u, v) was the last edge which caused the algorithm to set $C = C_u \cup C_v$. W.l.o.g. assume that $|C_u| \leq |C_v|$ (in case of ties let C_u be the connected component that is moved by the algorithm). Then set $C_k = C_u$ and set $\mathbb{C}_{k-1} = C_v$. Now repeat this procedure for \mathbb{C}_{k-1} in place of C to obtain C_{k-1} and \mathbb{C}_{k-2} . Continue this procedure until C_1 is of appropriate size.

Note that Properties 1 and 2 follow immediately from the above construction. Property 3 follows from the definition of small-to-large steps and the choice of C_u above. Property 4 is guaranteed by the stopping criterion of the above recursion. \square

Lemma 10 will be crucial for the proofs of many upcoming claims in this section. The lemma asserts that when a connected component C is currently assigned to the (say) right server but at the end it will be assigned to the left server, then it must contain relatively many vertices that were initially assigned to the right server.

LEMMA 10. *Let C be a connected component with $|C| \geq 4$. Suppose that C is currently assigned to server S_i and that C will be assigned to server S_{1-i} when the algorithm terminates. Then C contains at least $|C|/4$ vertices which were initially assigned to S_i .*

PROOF. Assume w.l.o.g. that C is currently assigned to the right server and it will be assigned to the left server when the algorithm terminates. We show that at least a $1/4$ -fraction of the vertices in C must be black. This implies the lemma.

Let (C_1, \dots, C_k) be a doubling decomposition of C which exists by Lemma 9. Observe that C must be assigned to the same server as $C_1 \cup C_2$ after they were merged and after the algorithm processed the majority vote for $C_1 \cup C_2$ (by Properties 3 and 4 of doubling decompositions). Thus, $C_1 \cup C_2$ had a black majority, i.e., it contains at least $|C_1 \cup C_2|/2$ black vertices. Since $|C_1 \cup C_2| \geq |C|/2$, C must contain at least $|C|/4$ black vertices. \square

Now we bound the augmentation that is used by Algorithm 2.

LEMMA 11. *The load of both servers is bounded by $n/2 + 4\Delta$. Hence, Algorithm 2 uses at most 4Δ augmentation.*

PROOF. Assume that at some point during the execution of the algorithm the (w.l.o.g.) right server contains more vertices than the left server. We bound the load of the right server.

Recall from Lemma 3 that $\Delta \leq n/4$. We start by considering the case where $\Delta = n/4$. In this case, even moving all n vertices to the right server only causes augmentation $n/2 = 2\Delta$.

Now consider the case where $\Delta < n/4$. Since $\Delta < n/4$, the initial assignment of S_1 must contain more vertices from either V_0 or V_1 . Thus, exactly one of the ground truth components V_0 and V_1 must have a black majority (as the algorithm colored all vertices initially assigned to S_1 black). We assume w.l.o.g. that V_1 has this black majority. This implies that V_1 has $n/2 - \Delta > n/4 > \Delta$ black vertices and V_0 has Δ black vertices. Further, as the algorithm proceeds, the vertices from V_1 must be moved to the right server.

The right server contains at each point a (potentially empty) set of vertices from V_0 and a (potentially empty) set of vertices from V_1 . For the latter set we use the trivial upper bound of $n/2$, while for the earlier set we give a bound of $\Delta/4$. The lemma follows.

Consider a component C which is on the right server and a subset of V_0 . By Lemma 10, C contains at least $|C|/4$ black vertices.

As there are only Δ black vertices in the ground truth component V_0 and each component $C \subseteq V_0$ on the right server has at least a $1/4$ -fraction of black vertices, it follows that all components on the right server which are subsets of V_0 can only contain 4Δ vertices. \square

Having derived the bound for the augmentation, our next goal is to show that the cost paid by the algorithm is bounded by $O(\alpha\Delta \log \Delta)$. We start by bounding the cost paid by the algorithm for each connected component.

The following lemma implies that the algorithm pays nothing for components in which all vertices have the same color.

LEMMA 12. *Let C be a connected component and suppose all vertices in C have the same color. Then the algorithm has never moved the vertices in C .*

PROOF. We prove the claim by induction over $s = |C|$.

Let $|C| = s = 1$. Then C consists of a single vertex. But the algorithm never moves single vertices unless they become part of a larger connected component. Hence, C is not moved.

Now let $|C| = s + 1$. Consider the last edge (u, v) which was inserted that forced the algorithm to merge $C = C_u \cup C_v$. Since in C all vertices have the same color, all vertices in C_u and C_v must have the same color. By induction hypothesis, the vertices in C_u and C_v have never been moved before. Thus, C_u and C_v must be assigned to the same server. This implies that a small-to-large step would not move C_u or C_v . Further, a majority voting step would not move $C_u \cup C_v$ since all vertices vote for the server which they are already assigned to. Thus, no vertices in C are moved. \square

Next, we bound the cost paid for any connected component.

LEMMA 13. *Let C be a connected component. Then the cost (over the entire execution time of the algorithm) paid for the vertices in C is at most $O(\alpha|C| \log |C|)$.*

PROOF. Consider a vertex $u \in C$. We perform the following accounting: we assign a token to u each time when it is reassigned to a server and we show that the number of tokens for u is bounded by $O(\log |C|)$. This implies that the total number of reassignments for the vertices in C is $O(|C| \log |C|)$ and the lemma follows.

First, consider the case where u is moved because it is in a smaller connected component (Line 7). Whenever this happens the size of the connected component containing u at least doubled. This can only happen $O(\log |C|)$ times.

Second, consider the case when u is moved because of a majority vote. A majority vote is performed every time when the size of the component containing u doubled. This can only happen $O(\log |C|)$ times and, hence, this can only add another $O(\log |C|)$ tokens for u .

Thus, the total number of tokens assigned to u is $O(\log |C|)$. \square

Note that Lemma 13 is only useful for components of size at most $O(\Delta)$: If we were to apply the lemma to a component C of size $\Theta(n)$ then the cost would only be bounded by $O(\alpha n \log n)$. However, this can be much worse than our desired bound of $O(\alpha \Delta \log \Delta)$ when $\Delta \ll n$. Thus, we need a more fine-grained argument to obtain our goal of showing that the cost paid by Algorithm 2 never exceeds $O(\alpha \Delta \log \Delta)$. To do this, we first prove two technical lemmas.

LEMMA 14. *Suppose C is a component which is moved from S_i to S_{1-i} and the vertices in C are never reassigned after this move. Then C contains at least $|C|/8$ vertices which were initially assigned to S_i .*

PROOF. There are only two possible reasons why C is moved: Either due to a small-to-large step (Line 7) or due to a majority voting step (Line 9). We consider both cases separately.

Case 1: C is moved due to a small-to-large step. Then by Lemma 10, C must contain at least $|C|/4$ vertices which were initially assigned to S_i .

Case 2: C is moved due to a majority voting step.

First, consider the case when C contains at most 7 vertices. Then at least one vertex was initially assigned to S_i (if all vertices had been initially assigned to S_{1-i} , they would all have the same color and a majority vote would not move C due to Lemma 12). Thus, at least a $1/7$ -fraction of the vertices were initially assigned to S_i and the lemma holds.

Second, suppose that C contains at least 8 vertices. Consider the last edge (u, v) that caused the merge $C = C_u \cup C_v$. Suppose that the small-to-large step moved C_u to the server of C_v . Note that C_v was assigned to S_i and C_u was moved to S_i . Now apply Lemma 10 to C_v . This implies that C_v must contain at least $|C_v|/4$ vertices that were initially assigned to S_i . As $|C_u| \leq |C_v|$, C must contain at least $|C|/8$ vertices that were initially assigned to S_i . \square

We are now ready to show that the cost incurred by the majority-voting algorithm never exceeds $O(\alpha\Delta \log \Delta)$.

LEMMA 15. *The total cost paid by Algorithm 2 is at most $O(\alpha\Delta \log \Delta)$ and the final assignment is a perfect partitioning.*

PROOF. When the algorithm finishes, the final assignment must be a perfect partitioning because the connected components were completely revealed. We only need to prove that the cost of the algorithm is $O(\alpha\Delta \log \Delta)$.

Recall that OPT moves exactly 2Δ vertices (Lemma 3). We can assume w.l.o.g. that OPT moves Δ vertices from V_0 that were initially assigned to S_1 to S_0 and Δ vertices from V_1 that were initially assigned to S_0 to S_1 . We will argue that the cost paid by the algorithm for moving all vertices from V_0 into the S_0 will be $O(\alpha\Delta \log \Delta)$; the same will hold for V_1 and S_1 symmetrically.

Consider time T during the execution of the algorithm where the following happens. A connected component C is reassigned the left server and C has the following properties: (1) C is a subset of V_0 and (2) the vertices in C never leave the left server after time T . Since each vertex of V_0 is assigned to the left server when the algorithm terminates, each vertex of V_0 is contained in a component with the above properties (when a vertex or component is never moved, we set $T = 0$). We call a component with the above properties *mixed* if it contains at least one black vertex. Note that when mixed component C is assigned to the left server, C contains a black vertex and, hence, C must be moved from the right to the left server.

We now bound the cost for mixed components. Let X be the set of all mixed components and let $C \in X$. Since C is mixed, Lemma 14 implies that at least $|C|/8$ vertices of C are black. As the black vertices in mixed components form a partition of the Δ black vertices in V_0 moved by OPT, we obtain that the number of black vertices in mixed components is Δ . Thus, the total number of vertices in all mixed components is $\sum_{C \in X} |C| \leq 8\Delta$.

By Lemma 13, the total cost paid for each $C \in X$ until (including) its final move is $O(\alpha|C| \log |C|)$. Since (by assumption) the vertices in C never move between the servers again, their cost never exceeds $O(\alpha|C| \log |C|)$ until the algorithm finishes. Hence, the cost paid by the algorithm for all mixed components is

$$\sum_{C \in X} O(\alpha|C| \log |C|) \leq \sum_{C \in X} O(\alpha|C| \log \Delta) = O(\alpha\Delta \log \Delta).$$

Now consider the vertices of V_0 which are not part of mixed components. These vertices must have been part of components in which all vertices are colored yellow. By Lemma 12, these vertices have never been moved. Thus, they do not incur any additional cost for the algorithm. \square

PROOF OF PROPOSITION 7. Lemma 11 gives the bound for the augmentation used by the algorithm. By Lemma 15 and Lemma 3, Algorithm 2 obtains a competitive ratio of

$$\frac{\text{ON}}{\text{OPT}} = \frac{O(\alpha\Delta \log \Delta)}{2\alpha\Delta} = O(\log \Delta) = O(\log n). \quad \square$$

3.4 Bringing It All Together: Theorem 2

PROOF. Proof of Theorem 2. Consider the following algorithm: Run the majority-voting algorithm until we have seen all edges or until at some point it tries to exceed the allowed augmentation. In the latter case, compute a perfectly balanced assignment respecting the connected components and start running Algorithm 1 (Section 3.2.2).

To prove the theorem, we distinguish two cases based on Δ .

First, suppose $\Delta < \varepsilon n/4$. By Proposition 7, Algorithm 2 uses at most 4Δ augmentation. Thus, in the current case the augmentation used by Algorithm 2 is bounded by $4\Delta < \varepsilon n$ and it is $O(\log n)$ -competitive. This proves the theorem for this case.

Second, suppose $\Delta \geq \varepsilon n/4$. In this case we run Algorithm 2 until it tries to exceed the allowed augmentation; this serves as a certificate that $\Delta \geq \varepsilon n/4$. At this point we switch to Algorithm 1.

When we switch algorithms, Algorithm 2 has paid $O(\alpha n \log n)$, by applying Lemma 13 to each connected component, and then summing over these costs. For switching to the perfectly balanced reassignment, we only need to pay $O(\alpha n)$ once.

By Proposition 6, Algorithm 1 never uses more than $O(n \log n + (\Delta \log n)/\varepsilon)$ vertex moves. Using the bound $\Delta \geq \varepsilon n/4$ and the fact that OPT pays $2\alpha\Delta$ (Lemma 3), we obtain the desired competitive ratio:

$$\frac{\text{ON}}{\text{OPT}} = O\left(\frac{\alpha n \log n + (\alpha \Delta \log n)/\varepsilon}{\alpha \Delta}\right) = O\left(\frac{\log n}{\varepsilon}\right). \quad \square$$

4 GENERALIZATION TO MANY SERVERS

We extend our study to the scenario with ℓ servers. As we will see, while several concepts introduced for the two server case are still useful, the ℓ -server case introduces additional challenges. We derive the following main result.

THEOREM 16. *Given a system with ℓ servers each of capacity $(1 + \varepsilon)n/\ell$ (i.e., augmentation $\varepsilon n/\ell$), for $\varepsilon \in (0, 1/2)$, then there exists an $O((\ell \log n \log \ell)/\varepsilon)$ -competitive algorithm.*

Our algorithm will be based on a recursive bipartitioning scheme, described in Section 4.1. We will use this bipartitioning scheme to derive a static approximation algorithm of the optimal solution (Section 4.2). Then we provide a recursive version of the majority voting algorithm which we will compare against the approximation algorithm (Section 4.3). In Section 4.4, we analyze the Small–Large–Rebalance algorithm in the ℓ server setting and we conclude by proving Theorem 16 in Section 4.5

4.1 The Bipartition Tree

We establish a recursive bipartitioning scheme of the servers which we will be using throughout the rest of this section. All algorithms in this section which use the recursive bipartitioning create such a bipartitioning at the start of the algorithm, before the adversary provides any edge. After that the bipartitioning will never be changed.

We obtain the bipartition scheme by growing a balanced binary tree on a set of ℓ leaves, where each leaf corresponds to a server S_i . We call this tree the *bipartition tree* and denote it by \mathcal{T} .

We denote the internal nodes of \mathcal{T} by w_1, \dots, w_s . For an internal node w_j , we write $T(w_j)$ to denote the subtree of T which is rooted at w_j and we define $S(w_j)$ to be the set of servers which are leaves in $T(w_j)$. We further write $V(w_i)$ to denote the set of vertices which are assigned to the servers in $S(w_j)$. See Figure 3 for an illustration.

Observe that \mathcal{T} defines a bipartition scheme: let w be an internal node of \mathcal{T} and let w_0, w_1 be its children. Then³ $S(w_0)$ and $S(w_1)$ are disjoint and their union is $S(w)$. Thus, \mathcal{T} implies a bipartition scheme of the servers and internal nodes correspond to bipartition steps.

Note that since \mathcal{T} is a balanced binary tree, there are $\ell - 1$ internal nodes in total and each server is contained in at most $\lceil \log \ell \rceil$ subtrees of T . Hence, for each server S_j there are at most $\lceil \log \ell \rceil$ internal vertices w such that $S_j \in S(w)$.

³If w_j is a leaf corresponding to server S , we set $S(w_j) = \{S\}$.

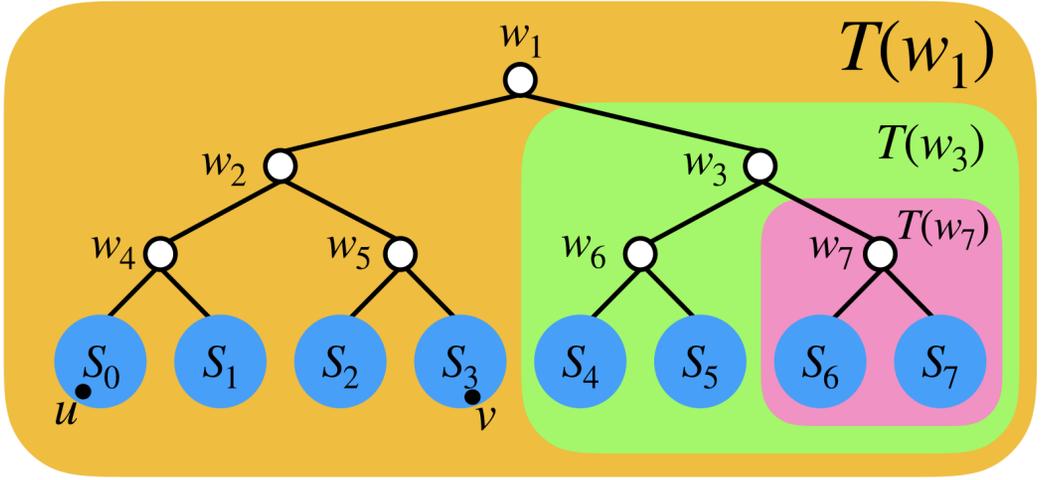


Fig. 3. An illustration of the bipartition tree \mathcal{T} for servers S_0, \dots, S_7 . The internal nodes of the bipartition tree are denoted w_1, \dots, w_7 . We highlighted the subtrees $\mathcal{T} = T(w_1)$, $T(w_3)$, and $T(w_7)$. Here we obtain the server sets $S(w_1) = \{S_0, \dots, S_7\}$, $S(w_3) = \{S_4, \dots, S_7\}$, and $S(w_7) = \{S_6, S_7\}$.

In the following we will refer to the internal nodes in \mathcal{T} as *nodes*, whereas the vertices V from the graph G are called *vertices*.

4.2 Offline Approximation Algorithm

We are not aware of a concise characterization of the optimal solution used by OPT (unlike in the two-server case in Section 3.1). Thus, to get a better understanding of the solution obtained by OPT, we provide an *offline* approximation algorithm, called APPROX, which exploits the previously defined bipartition scheme and which obtains a 2-approximation of the optimal solution. However, unlike the solution obtained by OPT, we allow the approximation algorithm to use *unlimited* augmentation in each server; its only goal is to move all vertices from the same ground truth components to the same server using few vertex moves.⁴ Later, APPROX will play a role for the design and analysis of our online algorithm.

Intuitively, APPROX traverses the bipartition tree \mathcal{T} top-down and greedily minimizes the number of vertices “moved over” each server bipartition. We now describe the algorithm in more detail.

APPROX is given the sequence of edges $\sigma = (e_1, \dots, e_r)$ *a priori* and it also knows the initial assignment $V_{\text{init}}(S_0), \dots, V_{\text{init}}(S_{\ell-1})$ of the vertices to the ℓ servers. Using the knowledge about the edges, APPROX starts by computing the connected components of G and obtaining the ground truth components $V_0, \dots, V_{\ell-1}$.

Now, for each ground truth component V_i , APPROX does the following. Let r be the root of \mathcal{T} and let w_0 and w_1 denote its children. Let $A_{ij} = V(w_j) \cap V_i$, $j = 0, 1$, denote the vertices from V_i which are currently assigned to servers in $S(w_j)$. Define $n_{ij} = |A_{ij}|$ and assume w.l.o.g. that $n_{i0} \geq n_{i1}$. The algorithm marks the vertices from A_{i1} as *dirty*. Now the algorithm recurses on the subtree $T(w_0)$ in place of \mathcal{T} and marks more vertices of V_i as dirty. The recursion stops when $S(w_0)$ only contains a single server S . Then the algorithm moves all dirty vertices of V_i into server S .

The pseudocode of APPROX is stated in Algorithm 3.

⁴In this setting, a trivial solution assigns all vertices to the same server at cost $O(\alpha n)$.

Algorithm 3 The static approximation algorithm APPROX

Input: All edges e_1, \dots, e_r at once

- 1: **procedure** APPROX(e_1, \dots, e_r)
- 2: Compute the connected components $V_0, \dots, V_{\ell-1}$ of G
- 3: **for** $i = 0, \dots, \ell - 1$ **do**
- 4: RECURSIVESTEP(\mathcal{T}, i)
- 5: **procedure** RECURSIVESTEP(T, i)
- 6: **if** T contains only a single server S **then**
- 7: Move all dirty vertices of V_i into S
- 8: **return**
- 9: Let r be the root of T and denote its children w_0, w_1
- 10: $A_{ij} \leftarrow V(w_j) \cap V_i, j = 0, 1$
- 11: $n_{ij} \leftarrow |A_{ij}|$ and suppose w.l.o.g. that $n_{i0} \geq n_{i1}$
- 12: Mark all vertices in A_{i1} dirty
- 13: RECURSIVESTEP($T(w_0), i$)

By overloading notation, we let APPROX denote the cost paid by APPROX. Further, we let APPROX_{*i*} denote the cost paid by APPROX to move all vertices from V_i to the same server S .

We now show that APPROX indeed yields a 2-approximate solution of the cost of the optimal offline algorithm.

LEMMA 17. APPROX $\leq 2 \cdot$ OPT.

PROOF. Fix any $i \in \{0, \dots, \ell - 1\}$. Let OPT_{*i*} denote the cost paid by OPT to move the vertices from V_i to the same server. We show that APPROX_{*i*} $\leq 2 \cdot$ OPT_{*i*}. This claim implies the lemma since

$$\text{APPROX} = \sum_i \text{APPROX}_i \leq 2 \sum_i \text{OPT}_i = 2 \cdot \text{OPT}.$$

Observe that while APPROX_{*i*} proceeds, it traverses \mathcal{T} from root r to one of the leaves, and at each step, it increases the level of the current internal node by one.

Using the solution of OPT_{*i*}, we can define a similar traversal of \mathcal{T} : Let r be the root of \mathcal{T} and let w_0 and w_1 be its children. As OPT_{*i*} must move all vertices from V_i to the same server S , OPT_{*i*} moves the vertices from $A_{ij} = V(w_j) \cap V_i$ to a server S in $S(w_{1-j})$ for $j \in \{0, 1\}$. We call the moved vertices *dirty*. After this move, OPT_{*i*} still needs to process the vertices of V_i which were initially assigned to a server in $S(w_{1-j})$ but not to S . We can view this as processing $T(w_{j-1})$. Thus, OPT_{*i*} traverses \mathcal{T} until the final server S is reached and marks a subset of V_i dirty.

The previous paragraphs define to two different traversals of \mathcal{T} and two different sets of dirty vertices. Let h be the smallest level where the two traversals picked different internal nodes in \mathcal{T} .

Until level $h - 1$, both vertices have marked the same W vertices dirty. At levels h and below, we obtain the following bounds. Let w be the internal node at level $h - 1$ that is traversed by both algorithms and let w_0, w_1 denote its children at level h . Let $n_{ij} = |V(w_j) \cap V_i|$ be defined as in the definition of APPROX. APPROX_{*i*} marks at most $|V(w)| = n_{i0} + n_{i1}$ vertices from $V(w)$ as dirty. Since the two traversals split at level h and APPROX_{*i*} moves $n_{i1} \leq n_{i0}$ vertices (by definition), OPT_{*i*} moves at least n_{i0} vertices.

Recall that for each algorithm, its sets of dirty vertices and its set of moved vertices are identical. Now the following computation proves the claim that APPROX_{*i*} $\leq 2 \cdot$ OPT_{*i*}:

$$\frac{\text{APPROX}_i}{\text{OPT}_i} \leq \frac{\alpha(W + n_{i0} + n_{i1})}{\alpha(W + n_{i0})} \leq \frac{W + 2n_{i0}}{W + n_{i0}} \leq 2. \quad \square$$

4.3 The Recursive Majority Voting Algorithm

We now describe an algorithm which works efficiently in the setting with ℓ servers whenever OPT does not perform too many vertex moves. The algorithm can be viewed as a generalization of Algorithm 2 to ℓ servers, by exploiting the previously defined bipartitioning scheme.

4.3.1 The Algorithm. The algorithm consists of two parts: A single *global algorithm* and multiple *local algorithms*, one per internal node in \mathcal{T} . The global algorithm maintains a recursive bipartitioning scheme (as defined in Section 4.1) and runs a local algorithm on each bipartition. The local algorithms are used to “reduce” the setting with multiple servers to the case with two servers.

We now describe the two parts in more detail and state the pseudocode in Algorithm 4. We write S_u to denote the server which vertex u is assigned to.

Global Algorithm. The global algorithm starts by computing the bipartition tree \mathcal{T} . On each internal node w of \mathcal{T} , the global algorithm instantiates a local algorithm which we describe below.

Furthermore, the global algorithm iterates over all vertices and does the following for each $v \in V$. The algorithm finds all internal nodes w_i such that $v \in V(w_i)$ and labels v with w_i . This labelling of the vertices only takes into account the initial assignment of the vertices and will never be changed throughout the running time of the algorithm. For example, if the vertices u and v in Figure 3 are assigned to servers S_0 and S_3 in the initial assignment, their labels will be $\{w_1, w_2, w_4\}$ and $\{w_1, w_2, w_5\}$, respectively.

When the adversary provides an edge (u, v) , the global algorithm does the following. It locates the servers S_u and S_v . If $S_u = S_v$, the algorithm merges the components and continues with the next edge. If $S_u \neq S_v$, the global algorithm finds the internal node w in \mathcal{T} which is the lowest common ancestor of S_u and S_v . (For example, in Figure 3 the lowest common ancestor for u and v is w_2 .) Then the global algorithm gives the edge (u, v) to the local algorithm corresponding to w .

Local Algorithms. A local algorithm is run on an internal node w of \mathcal{T} . Let w_0 and w_1 denote the children of w in \mathcal{T} . Note that each local algorithm corresponds to a bipartition step where the servers in $S(w)$ are partitioned into subsets $S(w_0)$ and $S(w_1)$.

An instance of the local algorithm only receives edges (u, v) from the global algorithm when (1) their endpoints are assigned to servers $S_u, S_v \in S(w)$ and (2) S_u and S_v are in different sets of the bipartition, i.e., $S_u \in S(w_j)$ and $S_v \in S(w_{1-j})$.

When the global algorithm provides an edge (u, v) with the above properties, the local algorithm locates C_u, C_v, S_u and S_v . Assume w.l.o.g. that $|C_u| \leq |C_v|$. Then C_u is moved to S_v and C_u and C_v are merged.⁵ As before, we call this a *small-to-large step*.

Finally, the local algorithm checks whether the new component $C_u \cup C_v$ has size n/ℓ or it surpassed a power of 2, i.e., it checks if $|C_u \cup C_v| = n/\ell$ or there exists an $i \in \mathbb{N}$ s.t. $|C_u| < 2^i$, $|C_v| < 2^i$ and $|C_u \cup C_v| \geq 2^i$. If this is the case, the local algorithm triggers a majority voting step for $C_u \cup C_v$ which we explain next.

Majority Voting Step. When a local algorithm triggers a majority voting step for a connected component C , the algorithm does the following. Let r be the root of \mathcal{T} and let w_0 and w_1 be the two child nodes of r . For $j \in \{0, 1\}$, let n_j denote the number of vertices in C with label w_j . If $n_j \geq n_{1-j}$, the algorithm recurses on w_j in place of w ; else, the algorithm recurses on w_{1-j} in place of w . The recursion continues until a leaf in the bipartitioning tree is reached which corresponds to a server S . Then the algorithm moves C to S .

⁵When C_u changes its server, all local algorithms corresponding to internal nodes w with $S_u \in S(w)$ or $S_v \in S(w)$, must be informed about this move. This can be done by recomputing $V(w)$ for each internal node w . Note that this is just an internal operation of the data structure and does not incur any cost to the algorithm.

Algorithm 4 The Recursive Majority Voting Algorithm

Input: A sequence of edges $\sigma = (e_1, \dots, e_r)$

- 1: **procedure** GLOBALALGORITHM(e_1, \dots, e_r)
- 2: Create a bipartition tree \mathcal{T} ▷ Initialization phase
- 3: **for** each internal node w of \mathcal{T} **do**
- 4: Instantiate LOCALALGORITHM(w)
- 5: **for** $v \in V$ **do**
- 6: Label v with each internal node w of \mathcal{T} s.t. $v \in V(w)$
- 7: **for** $i = 1, \dots, r$ ▷ Processing of the edges
- 8: $(u, v) \leftarrow e_i$
- 9: **if** $S_u = S_v$ **then**
- 10: Merge C_u and C_v , **continue**
- 11: $w \leftarrow$ the lowest common ancestor of S_u and S_v in \mathcal{T}
- 12: LOCALALGORITHM($w, (u, v)$)
- 13: **procedure** LOCALALGORITHM($w, (u, v)$)
- 14: $w_0, w_1 \leftarrow$ the children of w in \mathcal{T}
- 15: Suppose w.l.o.g. that $|C_u| \leq |C_v|$
- 16: Check if moving C_u to S_v triggers the stopping criterion
- 17: Move C_u to S_v and merge C_u and C_v ▷ Small-to-large step
- 18: **if** $|C_u \cup C_v| = n/\ell$ **or** there exists an $i \in \mathbb{N}$ s.t. $|C_u| < 2^i$, $|C_v| < 2^i$ and $|C_u \cup C_v| \geq 2^i$ **then**
- 19: MAJORITYVOTINGSTEP($C_u \cup C_v$)
- 20: **procedure** MAJORITYVOTINGSTEP(C)
- 21: $r \leftarrow$ the root of \mathcal{T}
- 22: $w_0, w_1 \leftarrow$ the children of r in \mathcal{T}
- 23: $n_j \leftarrow$ the number of vertices labeled with w_j in C , $j = 0, 1$
- 24: **if** $n_0 \geq n_1$ **then** $r \leftarrow w_0$ **else** $r \leftarrow w_1$
- 25: **if** $S(r)$ contains only one server **then**
- 26: Check if moving C to $S(r)$ triggers the stopping criterion
- 27: Assign C the single server in $S(r)$
- 28: **else** Go to Line 22

Note that the above majority voting procedure is very similar to what APPROX does for a single ground truth component V_i .

Stopping Criterion. To ensure that the algorithm does not exceed the augmentation of the servers, we add a stopping criterion.

To define the stopping criterion, let w be an internal node of \mathcal{T} with children w_0 and w_1 . For $j \in \{0, 1\}$, we call w_j *overloaded* if $V(w_j)$ contains at least $\epsilon n / (\ell \lceil \log \ell \rceil)$ vertices with label w_{1-j} .

Intuitively, the condition states that an internal node w_j is overloaded when its servers $S(w_j)$ obtained “many” vertices which were initially assigned to the other side of the bipartition, $S(w_{1-j})$.

The *stopping criterion* is checked before each component move (i.e., before each small-to-large step and before each majority voting step). It is triggered if the component move would create an assignment in which there exists an overloaded internal node w . When the stopping criterion is triggered, the global algorithm and all local algorithms stop and Algorithm 1 is started instead (we show in Section 4.4 that Algorithm 1 also works for ℓ servers).

4.3.2 Structural Properties. To obtain a better understanding of the algorithm, we first prove some structural properties about it and defer its cost analysis to Section 4.3.3. We consider the setting where each server has capacity $(1 + \varepsilon)n/\ell$ for $\varepsilon \in (0, 1/2)$.

In Subsections 4.3.2 and 4.3.3, we only analyze the cost Algorithm 4 without the cost of Algorithm 1. We analyze the cost of Algorithm 1 for ℓ servers in Sections 4.4 and 4.5.

We begin by showing that as long as the stopping criterion is not triggered, the vertex assignment created by Algorithm 4 is close to the initial assignment.

LEMMA 18. *Suppose the stopping criterion is not triggered. Then:*

- (1) *Each server contains at most $\varepsilon n/\ell$ vertices that were not initially assigned to it.*
- (2) *Each server contains at least $(1 - \varepsilon)n/\ell$ vertices that were initially assigned to it.*

PROOF. Consider any server S_j . We show that since the stopping criterion is not triggered, $V(S_j)$ obtains at most $\varepsilon n/(\ell \lceil \log \ell \rceil)$ vertices for each of the $\lceil \log \ell \rceil$ subtrees in \mathcal{T} containing S_j .

As argued in Section 4.1, there are at most $\lceil \log \ell \rceil$ internal nodes w of \mathcal{T} such that $S_j \in S(w)$. Since the stopping criterion is not triggered, no internal node of \mathcal{T} is overloaded.

To prove Part (1), consider an internal node w of \mathcal{T} with $S_j \in S(w)$. Let w_0, w_1 be the children of w and suppose $S_j \in S(w_r)$. Observe that $V(S_j)$ can obtain at most $\varepsilon n/(\ell \lceil \log \ell \rceil)$ vertices that were originally assigned to servers in $S(w_{1-r})$ (if it had received more vertices, then w_{1-r} would be overloaded). As there are at most $\lceil \log \ell \rceil$ nodes w with the above property, the number of vertices which were not initially assigned to S_j is bounded by $\varepsilon n/\ell$.

Now let us prove Part (2). Consider an internal node w of \mathcal{T} with $S_j \in S(w)$. Let w_0, w_1 be the children of w and suppose $S_j \in S(w_r)$. Now observe that the servers in $S(w_{1-r})$ can have obtained $\varepsilon n/(\ell \lceil \log \ell \rceil)$ vertices that were originally assigned to S_j (if they had received more vertices, then w_{1-r} would be overloaded). As there are at most $\lceil \log \ell \rceil$ nodes w with the above property, it follows that the number of vertices assigned to servers $\{S_0, \dots, S_{\ell-1}\} \setminus \{S_j\}$ that were initially assigned to S_j is $\varepsilon n/\ell$. Hence, S_j must contain at least $(1 - \varepsilon)n/\ell$ vertices that were initially assigned to it. \square

As a corollary of Lemma 18 we obtain the following lemma.

LEMMA 19. (1) *As long as the stopping criterion is not triggered, the load of each server is bounded by $(1 + \varepsilon)n/\ell$, i.e., Algorithm 4 uses only $\varepsilon n/\ell$ augmentation.*

(2) *When the stopping criterion is triggered, the augmentation still does not exceed $\varepsilon n/\ell$.*

PROOF. Part (1) of the lemma follows immediately from Part (1) of Lemma 18. Let us prove Part (2): The stopping criterion is checked every time before a component is moved. Hence, at the time when the algorithm checks the stopping criterion, the algorithm did not exceed the augmentation bound due to Part (1). If the algorithm triggers the stopping criterion, then the component was not yet moved and the augmentation is still the same as before. \square

Define the *final assignment* to be the assignment which is created by Algorithm 4 once it has seen all edges in G . We show that the final assignment of the algorithm provides a perfect partitioning if the stopping criterion is not triggered.

LEMMA 20. *If Algorithm 4 stops and the stopping criterion is not triggered, then the final assignment is a perfect partitioning.*

PROOF. By definition of the algorithm, vertices of the same connected component are always assigned to the same server. When the algorithm finishes, all edges of G were revealed and each component has size n/ℓ . By Lemma 18, the augmentation of each server is at most $\varepsilon n/\ell$. Since $\varepsilon < 1/2$, no server can have more than one component assigned. As each component is placed on a

server, each component is placed alone on a server. This proves that the algorithm creates a perfect partitioning. \square

Indeed, we show that the final assignment of Algorithm 4 is not only a perfect partitioning, but it is the same assignment as the one created by APPROX from Section 4.2.

LEMMA 21. *If Algorithm 4 stops and the stopping criterion is not triggered, Algorithm 4 and APPROX have the same final assignment.*

PROOF. By Part (2) of Lemma 18, Algorithm 4 moves at most $\varepsilon n/\ell$ vertices out of each server compared to the initial assignment. Hence, in the final assignment each server must still contain at least $(1 - \varepsilon)n/\ell > n/(2\ell)$ vertices from its original assignment since $\varepsilon \in (0, 1/2)$. Thus, in the final assignment each server contains more than half of the vertices that were originally assigned to it.

Consider any server S_j and let $V_{\text{init}}(S_j)$ be the set of vertices initially assigned to S_j . Then there must exist a ground truth component V_i with $|V_i \cap V_{\text{init}}(S_j)| \geq n/(2\ell)$. We show that APPROX and Algorithm 4 both assign this component V_i to S_j . This proves the lemma since this claim holds for any S_j .

First, consider APPROX. Note that at each step of the traversal of \mathcal{T} , the majority of the vertices in V_i will vote for the internal node containing server S_j . Hence, APPROX will place V_i on S_j .

Second, consider Algorithm 4. When the algorithm stops, all edges were revealed and the connected components agree with the ground truth components. Now consider the component $C = V_i$. When the C grows to size $|C| = n/\ell$, the algorithm performs a majority voting step (by definition of the algorithm). At this point, more than half of the vertices in C were labeled with S_j (because more than half of the vertices from $C = V_i$ were originally assigned to S_j). Hence, Algorithm 4 will also place V_i on S_j . \square

4.3.3 Analysis. The rest of this subsection is devoted to proving the following proposition about Algorithm 4.

PROPOSITION 22. *Suppose there are ℓ servers and each has capacity $(1 + \varepsilon)n/\ell$ for $\varepsilon \in (0, 1/2)$, i.e., the augmentation is $\varepsilon n/\ell$. Algorithm 4 has the following properties:*

- (1) *If the stopping criterion is not triggered, the algorithm creates a perfect partitioning, its cost is bounded by $O(\text{OPT} \cdot \log n)$ and at no point during its execution it uses more than $\varepsilon n/\ell$ augmentation.*
- (2) *If the stopping criterion is triggered, the cost of the algorithm is $O(\alpha n \log n)$ plus the cost of Algorithm 1 and the cost of OPT is at least $\Omega(\alpha \varepsilon n/(\ell \log \ell))$.*

We prove the proposition at the end of this section. We start by proving a sequence of lemmata and begin by reasoning about the cost paid by Algorithm 4. As shown in Lemma 1 we only need to bound the moving cost paid by Algorithm 4 to bound its total cost.

The following lemma bounds the cost paid for any connected component C .

LEMMA 23. *Let C be a connected component. Then the cost (over the entire execution time of the algorithm) paid for moving the vertices in C is $O(\alpha |C| \log |C|)$.*

PROOF. We can use the same accounting argument as in the proof of Lemma 13. That is, we assign a token to a vertex v whenever it is moved. Now, whenever the component C containing v is moved due to a small-to-large step, the size of C doubles. This can only happen $O(\log |C|)$ times. Furthermore, there are only $O(\log |C|)$ majority voting steps involving u : Each majority voting step is triggered because $|C| = n/\ell$ or because $|C|$ surpassed a power of 2; the first event can happen only once and the second event can happen at most $O(\log |C|)$ times. Hence, v will never

accumulate more than $O(\log |C|)$ tokens. Since the above arguments apply for each $v \in C$, the total cost paid for moving the vertices in C is bounded by $O(|C| \log |C|)$. \square

Let $f: V \rightarrow \{0, \dots, \ell - 1\}$ be the function which maps each vertex to its server in the final assignment by Algorithm 4. That is, when Algorithm 4 processed all edges, each v is assigned to $S_{f(v)}$. For a connected component C , set $f(C) = f(u)$ for $u \in V$. Note that $f(C)$ is well-defined since all vertices of C are assigned to the same $S_{f(C)}$ when the algorithm terminates.

In the following proofs, we will write $\#w(C)$ to denote the number of vertices in a connected component C which are labeled with w . We further write $\overline{\#w(C)}$ to denote the number of vertices in C which are *not* labeled with w , i.e., $\overline{\#w(C)} = |C| - \#w(C)$.

Lemma 24 shows that whenever a component C is assigned to a server which is not its final server, it must contain relatively many vertices which were not initially assigned to its final server $S_{f(C)}$.

LEMMA 24. *Consider any point in the execution of the algorithm at which a connected component C is assigned to server $S \neq S_{f(C)}$. Let w be the lowest common ancestor of S and $S_{f(C)}$ in \mathcal{T} and denote the children of w by w_0 and w_1 .*

If $S_{f(C)} \in S(w_j)$ for $j \in \{0, 1\}$, then:

- (1) *C contains at least $|C|/4$ vertices which do not have label w_j , i.e., $\overline{\#w_j(C)} \geq |C|/4$.*
- (2) *C contains at least $\overline{\#w_j(C)}$ vertices which were not initially assigned to $S_{f(C)}$.*

PROOF. To prove Part (1), consider a doubling decomposition (C_1, \dots, C_k) of C (see Definition 8); the decomposition exists by Lemma 9 which also applies in the ℓ server setting. After C_1 and C_2 were merged, Algorithm 4 performed a majority voting step and placed $C_1 \cup C_2$ in a server $S \in S(w_{1-j})$. Thus, $\#w_j(C_1 \cup C_2) \leq |C_1 \cup C_2|/2$ (otherwise, the majority voting step would have chosen a server in $S(w_j)$). Since $|C_1 \cup C_2| \geq |C|/2$ and $C_1 \cup C_2 \subseteq C$,

$$\begin{aligned} \overline{\#w_j(C)} &\geq \overline{\#w_j(C_1 \cup C_2)} = |C_1 \cup C_2| - \#w_j(C_1 \cup C_2) \\ &\geq |C_1 \cup C_2| - |C_1 \cup C_2|/2 = |C_1 \cup C_2|/2 \geq |C|/4. \end{aligned}$$

For Part (2) note that each vertex which was initially assigned to $S_{f(C)}$ has label w_j (because $S_{f(C)} \in S(w_j)$ by assumption). \square

In the following, we show that the cost paid by the algorithm is $O(\text{OPT} \cdot \log n)$ when the stopping criterion is not triggered. We start by showing that when a component is moved for the last time, it contains a large number of vertices which did not originate from the server it is assigned to.

LEMMA 25. *Let C be a component which is moved to server $S_{f(C)}$ and suppose the vertices of C are never reassigned after this move.⁶ Then C contains at least $|C|/8$ vertices which were not assigned to $S_{f(C)}$ in the initial assignment.*

PROOF. Note that C is moved due to one of two reasons: Either because of a small-to-large step or because of a majority voting step. We distinguish between these cases.

In case of a small-to-large step, C is assigned to a server $S \neq S_{f(C)}$ before the move. Lemma 24 implies that C contains at least $|C|/4$ vertices which were not originally assigned to $S_{f(C)}$.

Now suppose that C is moved due to a majority voting step. Let (u, v) be the last edge which was inserted and which triggered the majority voting step for C . Then Algorithm 4 previously merged components C_u and C_v ; suppose w.l.o.g. that C_u was moved to C_v and $|C_u| \leq |C_v|$. Prior to the majority voting step, C is assigned to the same server $S \neq S_{f(C)}$ that C_v was assigned to

⁶Note that when a small-to-large step is performed, two components are merged due to the corresponding edge insertion. In this case, the component C in the lemma is the component which is being moved (i.e., before merging).

before (u, v) was inserted. Hence, we can apply Lemma 24 to C_v and obtain that C_v contains at least $|C_v|/4$ vertices which were not initially assigned to $S_{f(C)}$. Thus, the number of vertices in C which do not originate from $S_{f(C)}$ is at least

$$|C_v|/4 \geq 2|C_v|/8 \geq |C_u \cup C_v|/8 = |C|/8. \quad \square$$

The next lemma considers the cost paid by Algorithm 4 when the stopping criterion is not triggered.

LEMMA 26. *Suppose there are ℓ servers and each has capacity $(1 + \varepsilon)n/\ell$ for $\varepsilon \in (0, 1/2)$, i.e., the augmentation is $\varepsilon n/\ell$. If the stopping criterion is not triggered and Algorithm 4 stops, then the cost paid by the algorithm is $O(\text{OPT} \cdot \log n)$.*

PROOF. Fix some $i \in \{0, \dots, \ell - 1\}$. Recall that APPROX_i denotes the cost paid by APPROX to move the vertices from V_i to the server $S_{f(V_i)}$. We show that for V_i , Algorithm 4 pays $O(\text{APPROX}_i \log n)$. The lemma follows from this claim and Lemma 17, since the total cost paid by Algorithm 4 is bounded by

$$\sum_i O(\text{APPROX}_i \cdot \log n) = O(\text{APPROX} \cdot \log n) = O(\text{OPT} \cdot \log n).$$

Consider any ground truth component V_i and let Δ denote the number of vertices APPROX_i moves to server $S_{f(C)}$. Note that as APPROX_i moves Δ vertices into $S_{f(C)}$, we get $\text{APPROX}_i = \alpha\Delta$.

Consider time T of the execution of the algorithm where the following happens. A component C is reassigned to $S_{f(C)}$ and C has the following properties: (1) C is a subset of V_i and (2) the vertices in C never leave server $S_{f(C)}$ after time T . Since each vertex of V_i is assigned to $S_{f(C)}$ when the algorithm terminates, each vertex of V_i is contained in a component with the above properties (when a vertex or component is never moved, we set $T = 0$). A component C with the above properties is a *mixed* component if C contains at least one vertex which was not initially assigned to $S_{f(C)}$. Note that when a mixed component C is reassigned to $S_{f(C)}$, C contains at least one vertex which was not initially assigned to $S_{f(C)}$ and, hence, C must be moved from a server S_y , $y \neq f(C)$, to $S_{f(C)}$.

We bound the cost for mixed components. Let X be the set of all mixed components of V_i . Recall that Algorithm 4 and APPROX create the same final assignment (Lemma 21). Hence, Algorithm 4 moves the same Δ vertices from V_i into $S_{f(V_i)}$ as APPROX. Lemma 25 implies that for each $C \in X$ at least $|C|/8$ vertices from C are part of the Δ vertices moved by APPROX. Thus, the union of all $C \in X$ contains at most 8Δ vertices.

By Lemma 23, Algorithm 4 pays at most $O(\alpha|C| \log |C|)$ for each $C \in X$ over the entire execution. Thus, its total cost is bounded by

$$\sum_{C \in X} O(\alpha|C| \log |C|) \leq O(\alpha\Delta \log n) = O(\text{ON}_i \cdot \log n).$$

Consider the vertices of V_i which are not in mixed components. These vertices must have been part of components in which all vertices were originally assigned to $S_{f(V_i)}$. By Lemma 12 (which still applies in the ℓ server setting), these vertices were never moved. Thus, they do not incur any cost to the algorithm. \square

Next, we show that when the stopping criterion is triggered, the recursive majority voting algorithm pays $O(n \log n)$ and cost of the solution by OPT is $\Omega(\alpha\varepsilon n/(\ell \log \ell))$.

LEMMA 27. *When the stopping criterion is triggered, (1) the cost paid by Algorithm 4 is $O(\alpha n \log n)$ and (2) the cost paid by OPT is $\Omega(\alpha\varepsilon n/(\ell \log \ell))$.*

PROOF. Let Y denote the set of all connected components. Part (1) follows from Lemma 23 since the total cost paid by Algorithm 4 is

$$\sum_{C \in Y} O(\alpha|C| \log |C|) \leq \sum_{C \in Y} O(\alpha|C| \log n) = O(\alpha n \log n).$$

Now we prove Part (2). Let w be an internal node of \mathcal{T} with children w_0, w_1 and suppose w.l.o.g. that w_0 is overloaded. Since the stopping criterion is triggered, $V(w_0)$ contains at least $\epsilon n / (\ell \log \ell)$ vertices with label w_1 .

Let X be the set of all connected components C with the following properties: C is assigned to a server in $S(w_0)$ at the time at which the stopping criterion is triggered and C contains at least one vertex which is labeled with w_1 .

To show that OPT performs $\Omega(\epsilon n / (\ell \log \ell))$ vertex moves, we prove that OPT performs $\Omega(\#w_1(C))$ vertex moves for each $C \in X$. Part (2) of the lemma follows since the components in X contain at least $\epsilon n / (\ell \log \ell)$ vertices with label w_1 and thus

$$\text{OPT} \geq \sum_{C \in X} \Omega(\alpha \#w_1(C)) = \Omega(\alpha \epsilon n / (\ell \log \ell)).$$

We prove that OPT moves at least $\Omega(\#w_1(C))$ vertices for each $C \in X$ by distinguishing two cases for $C \in X$. We define g as the function which maps $C \in X$ to the server it is assigned to in the solution of OPT, i.e., OPT assigns $C \in X$ to server $S_{g(C)}$.

Case 1: $S_{g(C)} \notin S(w_1)$, i.e., in the final assignment of OPT, the vertices in C are assigned to $S_{g(C)} \notin S(w_1)$. Then OPT must perform at least $\#w_1(C)$ vertex moves because it must move all w_1 -labeled vertices of C from their initial server in $S(w_1)$ to $S_{g(C)} \notin S(w_1)$.

Case 2: $S_{g(C)} \in S(w_1)$, i.e., in the final solution by OPT, the vertices in C are assigned to a server $S_{g(C)} \in S(w_1)$. We show that C contains at least $|C|/4$ vertices without label w_1 . This implies the claim since OPT must move at least $\overline{\#w_1(C)} \geq |C|/4$ vertices from servers not in $S(w_1)$ to $S_{g(C)} \in S(w_1)$.

Consider a doubling decomposition (C_1, \dots, C_k) of C (which exists by Lemma 9). After C_1 and C_2 were merged, the algorithm performed a majority voting step and placed $C_1 \cup C_2$ in a server in $S(w_0)$. Thus, $\#w_1(C_1 \cup C_2) \leq |C_1 \cup C_2|/2$ (otherwise, the majority voting step would place $C_1 \cup C_2$ in a server in $S(w_1)$). Hence, $\#w_1(C_1 \cup C_2) = |C_1 \cup C_2| - \#w_1(C_1 \cup C_2) \geq |C_1 \cup C_2|/2$. Since $|C_1 \cup C_2| \geq |C|/2$, we get $\#w_1(C) \geq |C|/4$. \square

PROOF OF PROPOSITION 22. The first statement of the proposition is implied by Lemmas 20 (perfect partitioning), 26 (total cost) and 19 (small augmentation). The second statement is proved in Lemma 27 (guarantees when stopping criterion is triggered). \square

4.4 Small–Large–Rebalance Algorithm for Many Servers

To obtain an efficient algorithm in cases where OPT moves many vertices, we reuse the Algorithm 1 from Section 3.2.2. Note that Algorithm 1 also works with ℓ servers because it did not use the fact that there are only two servers. In the setting with ℓ servers, we obtain the following result.

PROPOSITION 28. *Suppose that all servers have capacity $(1 + \epsilon)n/\ell$ for $\epsilon > 0$, i.e., the augmentation is $\epsilon n/\ell$. Then the cost paid by the more efficient version of Algorithm 1 is $O(\alpha n \log n + (\text{OPT} \cdot \ell \log n)/\epsilon)$.*

PROOF. The proof of the lemma is almost the same as the proof of Proposition 6. The only difference is that we need to bound the number of rebalance operations differently.

The number of vertex moves performed by the algorithm which always moves the smaller connected component to the server of the larger connected component is $O(n \log n)$ and, hence, it incurs cost $O(\alpha n \log n)$. Now, whenever a server exceeds its capacity, the algorithm must have

moved at least $\varepsilon n/\ell$ vertices. This can only happen $O(\ell \log n/\varepsilon)$ times. By the same arguments as in the proof of Proposition 6, each rebalancing operations costs $O(\text{OPT})$. Hence, the cost for all rebalancing steps is bounded by $O(\text{OPT} \cdot \ell \log n/\varepsilon)$. \square

We should point out that as in Lemma 5, we could also do the repartitioning step of Algorithm 1 by taking *any* perfectly balanced assignment respecting the connected components. In the analysis this would incur $\Theta(n)$ vertex moves for each such step and, hence, yield an algorithm with $O((n\ell \log n)/\varepsilon)$ vertex moves in total. However, unlike in the two-server case, finding a perfectly balanced assignment respecting the connected components is an NP-hard problem. Nonetheless, the problem can be solved approximately in polynomial time at the cost of a constant factor in the competitive ratio. We discuss this in further detail in Section 5.2.2.

4.5 Bringing It All Together: Theorem 16

PROOF OF THEOREM 16. Consider the algorithm which first runs Algorithm 4 until the stopping criterion is triggered and then switches to the Algorithm 1 from Section 4.4.

If the stopping criterion of the Algorithm 4 is not triggered, then by Proposition 22 the cost of the algorithm is $O(\text{OPT} \cdot \log n)$. Thus, it is $O(\log n)$ -competitive.

If the stopping criterion is triggered, then Algorithm 4 pays $O(\alpha n \log n)$ by Proposition 22 and the cost of OPT is $\Omega(\varepsilon n/(\ell \log \ell))$. Furthermore, the cost of Algorithm 1 is $O(\alpha n \log n + (\text{OPT} \cdot \ell \log n)/\varepsilon)$ by Proposition 28. Hence, we obtain the following competitive ratio:

$$\begin{aligned} \frac{O(\alpha n \log n + (\text{OPT} \cdot \ell \log n)/\varepsilon)}{\text{OPT}} &= \frac{O(\alpha n \log n)}{\text{OPT}} + O\left(\frac{\ell \log n}{\varepsilon}\right) \\ &\leq O\left(\frac{\alpha n \log n \cdot \ell \log \ell}{\alpha \varepsilon n}\right) + O\left(\frac{\ell \log n}{\varepsilon}\right) = O\left(\frac{\ell \log n \log \ell}{\varepsilon}\right). \end{aligned} \quad \square$$

5 DISTRIBUTED AND FAST ALGORITHMS

In this section we show how the algorithms from Section 4 can be implemented in a distributed setting (Section 5.1) and how they need to be modified to work in polynomial time at the cost of a slightly worse competitive ratio (Section 5.2).

We should point out that even though we discuss the distributed and polynomial time versions of the algorithms separately, they can easily be combined to obtain a distributed algorithm with polynomial computation time.

5.1 Distributed Algorithm

While in Section 4 we presented algorithms in a centralized model of computation, we now show how Algorithms 1 and 4 can be implemented in a distributed model of computation. For realistic parameter settings, the network traffic caused by our distributed algorithms does not increase (asymptotically) compared to the traffic caused by moving around the vertices between the servers.

In our distributed model of computation we assume that all servers have access to: (1) the number of servers ℓ , (2) the ID of the root server S_0 , (3) a shared clock, and (4) all-to-all communication.

When computing the network traffic, we will asymptotically count the number of messages sent by the algorithms and we further assume that each message contains $\Theta(\log n)$ bits. For the sake of simplicity we assume that moving a vertex from one server to another incurs cost $\alpha = \Theta(\log n)$.⁷ Because of this simplifying assumption we do not have to distinguish between the number of messages sent by the algorithm and the number of messages used for moving algorithms.

⁷Note that this is a realistic assumption since in order to move a vertex, a server must send the ID of a vertex to another server. Sending the ID of the vertex requires $\Theta(\log n)$ bits.

In this distributed model of computation, we obtain the following main result for the distributed versions of the algorithms.

THEOREM 29. *Consider a system with ℓ servers each of capacity $(1 + \epsilon)n/\ell$ (i.e., augmentation $\epsilon n/\ell$) for $\epsilon \in (0, 1/2)$. Let M be the number of vertex moves performed by *OPT*.*

Then there exists a distributed $O((\ell \log n \log \ell)/\epsilon)$ -competitive algorithm which sends

- (1) $O(M \log n)$ messages if $M = O(\epsilon n/(\ell \log \ell))$,
- (2) $O((\ell^2 \log n)/\epsilon + n \log n + (\text{OPT} \cdot \ell \log n)/\epsilon)$ messages if $M = \Omega(\epsilon n/(\ell \log \ell))$.

In particular, if $\ell = O(\sqrt{\epsilon n})$, then the algorithm's communication cost does not exceed its cost for moving vertices.

We show for Algorithm 4 (Section 5.1.1) and for Algorithm 1 (Section 5.1.2) individually how they can be implemented distributedly. After that we prove Theorem 29 in Section 5.1.3.

5.1.1 Making Algorithm 4 Distributed. We start by considering the distributed implementation of Algorithm 4 and obtain the following result.

LEMMA 30. *Algorithm 4 can be implemented in a distributed model of computation such that the guarantees from Proposition 22 still hold. Furthermore, if *OPT* performs M vertex moves, then we additionally have the following two properties:*

- (1) *If the stopping criterion is not triggered and the algorithm terminates, then the algorithm sent $O(M \log n)$ messages.*
- (2) *If the stopping criterion was triggered, the algorithm sent $O(n \log n)$ messages.*

PROOF. We start by presenting the necessary modifications to the algorithm and analyze the number of sent messages at the end of the proof.

Let us start by observing that each server can maintain a local representation of the bipartition tree \mathcal{T} : Since the number of servers ℓ is known to all servers and \mathcal{T} does not depend on any other quantity, each server can compute \mathcal{T} locally. Next, the data structure stores for each vertex its ID (requiring $O(\log n)$ bits) and the ID j of the server S_j it was initially assigned to (requiring $O(\log \ell)$ bits). Thus, the data structure uses $O(\log n)$ bits of storage for each vertex. In other words, it takes $O(1)$ messages to move a vertex between different servers.

Next, we provide the modifications for checking the stopping criterion, small-to-large steps and for majority voting steps.

Before the algorithm moves a component C from server S to server S' , S and S' need to check whether the move would trigger the stopping criterion. To do so, S and S' do the following. First, S asks S' for its ID using $O(1)$ messages. Second, S distinguishes between two cases: (1) C contains at most $\lceil \log \ell \rceil$ vertices. Then for each vertex $v \in C$, S sends a message to S' containing the ID of the server v was initially assigned to. This requires $O(|C|)$ messages. (2) C contains more than $\lceil \log \ell \rceil$ vertices. Then S locally computes all internal nodes w of the bipartition tree \mathcal{T} which contain S' as a leaf. For each such node w , let \bar{w} be the sibling of w in \mathcal{T} . Now for each w , S computes the number of vertices in C which were initially assigned to a server in $S(\bar{w})$. Then S sends these values to S' using $O(\log \ell)$ messages. Note that in both cases the algorithm does not send more than $O(|C|)$ messages and these messages can be charged to the moving cost of C (which requires $\Omega(|C|)$ messages) which happens after the checking of the stopping criterion. Third, S' receives the messages from S and checks locally whether receiving C would trigger the stopping criterion. If the stopping criterion is not triggered, S' tells S to start moving C . If the stopping criterion is triggered, S sends a message to the root server S_0 about this event. Then S_0 informs all other servers about switching to Algorithm 1. This requires $O(\ell) = O(n \log n)$ messages.

Now suppose the algorithm performs a small-to-large step and the stopping criterion was previously checked and not triggered. In this case, no modifications are necessary: The component C can just be sent from one server to the other at the cost of $O(|C|)$ messages (since each vertex in C can be sent using $O(1)$ messages).

Now suppose a server S needs to perform a majority voting step for a component C . First, observe that S can locally decide whether a majority voting step is necessary for C since it must only check the size of C . Second, when a majority voting step is necessary, S can locally compute which server S' will be the recipient of C : For each vertex $v \in C$, S knows which server v was initially assigned to. Hence, for each v , S can compute the labels of v w.r.t. the bipartitioning scheme from Section 4.1 locally. Since S also knows \mathcal{T} , S can compute to which server S' the component C should be moved to. These operations do not require any communication between the servers.

To conclude the proof of the lemma, observe that the distributed algorithm performs exactly as many vertex moves as the centralized algorithm. Hence, the guarantees from Proposition 22 still hold. Next, we analyze the number of messages sent by the algorithm. A small-to-large step moving a component C requires $O(|C|)$ messages. Checking the stopping criterion before moving a component C requires another $O(|C|)$ messages. Checking whether a majority voting step is necessary requires no communication at all. Hence, the number of messages used by the algorithm is linear in its number of vertex moves. Thus, Proposition 22 implies the two additional properties which are claimed in the statement of the lemma. \square

5.1.2 *Making Algorithm 1 Distributed.* For the distributed implementation of Algorithm 1 we obtain the following result.

LEMMA 31. *Algorithm 1 can be implemented in a distributed model of computation such that the guarantees from Proposition 28 still hold. Furthermore, if OPT performs M vertex moves, then the algorithm sends at most $O((\ell^2 \log n)/\epsilon + n \log n + (M\ell \log n)/\epsilon)$ messages.*

PROOF. We start by stating which modifications need to be made to make Algorithm 1 distributed.

First, suppose that Algorithm 1 performs a small-to-large step moving a component C and that this move does not make any server exceed its capacity. In this case, no modifications are necessary and the number of messages sent is $O(|C|)$ as we have seen in the proof of Lemma 30.

Second, suppose that a small-to-large step wants to move component C to server S which would cause S to exceed its capacity. Then the algorithm performs the following operations:

- (1) S informs the root server S_0 that a rebuild is required.
- (2) S_0 asks all ℓ servers to send the edges that were inserted and caused the merge of two connected components since the last rebuild. The servers send all these edges together with the timestamps when they were inserted.
- (3) S_0 locally simulates the whole system from the beginning and obtains knowledge about all connected components and which servers they are assigned to.
- (4) S_0 tells all other servers S_j which components need to be moved and all servers perform the necessary moves.

Since the distributed algorithm performs exactly the same vertex moves as the centralized algorithm, the distributed algorithm is correct and provides the same guarantees as provided in Proposition 28. We only need to analyze how many messages the algorithm sends. To do so, we analyze each step separately.

Every time Step 1 is performed, it requires $O(1)$ messages. As there are $O((\ell \log n)/\epsilon)$ rebuilds in total, Step 1 sends $O((\ell \log n)/\epsilon)$ messages in total.

To bound the number of messages sent in Step 2, recall that in total there are only $O(n)$ edges which merge connected components. Hence, sending these edges requires $O(n)$ messages. Furthermore, when a server did not obtain an edge merging two connected components between two rebuilds, it can inform S_0 about this in $O(1)$ messages. As this can be the case for at most ℓ servers and since there are $O((\ell \log n)/\varepsilon)$ rebuilds, at most $O((\ell^2 \log n)/\varepsilon)$ messages are sent when servers did not receive new edges.

In Step 3, S_0 locally simulates the system. This does not incur any network traffic.

Now consider Step 4. During a rebuild, the number of components which the algorithm needs to reassign is trivially bounded by the number of vertex moves performed during the rebuild. Thus, Proposition 28 implies that only $O(n \log n + (\text{OPT} \cdot \ell \log n)/\varepsilon)$ messages are required for all invocations of Step 4.

In total, we obtain that the algorithm sends at most $O((\ell^2 \log n)/\varepsilon) + n \log n + (M\ell \log n)/\varepsilon$ messages, where M is the number of vertices moved by OPT. \square

5.1.3 Proof of Theorem 29. To prove the claim about the competitive ratio of the algorithm observe that the distributed algorithm performs exactly the same vertex moves as the centralized algorithm. Hence, the cost paid by both algorithms is the same and the distributed algorithm has the same competitive ratio as the centralized algorithm in Theorem 16.

The claim about the number of messages sent by the algorithm follows from Lemma 30 and Lemma 31 and summing over the number of messages.

To prove the last claim of the theorem, we distinguish two cases. If the stopping criterion was not triggered, then the claim holds by Lemma 30. If the stopping criterion was triggered, then if $\ell = O(\sqrt{\varepsilon n})$, we obtain that the total number of messages is

$$\begin{aligned} O((\ell^2 \log n)/\varepsilon) + n \log n + (M\ell \log n)/\varepsilon &= O((\varepsilon n \log n)/\varepsilon + n \log n + (M\ell \log n)/\varepsilon) \\ &= O(n \log n + (M\ell \log n)/\varepsilon), \end{aligned}$$

which is exactly the number of vertices moved by Algorithm 1. \square

5.2 Fast Algorithms

In this section, we discuss the computational challenges when computing perfectly balanced assignments. These computational problems occur when Algorithm 1 performs rebalancing steps (see Section 3.2 and Section 4.4). So far, we were only concerned with algorithms which try to minimize the vertex moves while using potentially exponential running time. We now consider polynomial time algorithms. The only step where our algorithms might use exponential time is during rebalancing. Thus we show next how to perform the rebalancing operations in polynomial time. In the case of $\ell > 2$ servers, our polynomial time algorithms perform slightly more vertex moves than the exponential time algorithms.

We discuss the two server case which can be solved optimally in polynomial time in Section 5.2.1. In Section 5.2.2, we argue that in the general case with $\ell > 2$ servers this problem is NP-hard. We resolve this issue in Section 5.2.3 by computing approximately balanced assignments in polynomial time.

5.2.1 Computing Perfectly Balanced Assignments for Two Servers. We consider computing a perfectly balanced assignment respecting the connected components for two servers. Specifically, we provide a dynamic program which can find such an assignment in polynomial time.

The dynamic program works as follows. Suppose C_1, \dots, C_q are the connected components assigned to the two servers. Now let $k_i = |C_i|$ for $i = 1, \dots, q$. We create a set \mathcal{S} consisting of integers with the following property: Each number $s \in \mathcal{S}$ corresponds to a set of connected components \mathbb{C} such that $|\bigcup_{C \in \mathbb{C}} C| = s$. That is, whenever $s \in \mathcal{S}$, there exists a set of connected

components which together contain s vertices. For each $s \in \mathcal{S}$, the algorithm maintains a set of connected components explicitly. We denote the components corresponding to value $s \in \mathcal{S}$ by $\text{components}(s)$.

At the beginning of a rebalancing step, the algorithm sets $\mathcal{S} = \{0\}$. The connected component corresponding to value 0 is simply the empty set of vertices, i.e., $\text{components}(0) = \emptyset$. For $i = 1, \dots, q$ the algorithm does the following. Iterate over all $s \in \mathcal{S}$ and over all components and add $s + k_i$ to \mathcal{S} if $s + k_i \notin \mathcal{S}$ and $C_i \notin \text{components}(s)$. Whenever a new value $s + k_i$ is added to \mathcal{S} , set $\text{components}(s + k_i) = \text{components}(s) \cup \{C_i\}$.

As soon as the value $n/2$ is added to \mathcal{S} , the dynamic program stops and assigns all vertices in $\text{components}(n/2)$ to the left server and all remaining vertices to the right server.

The correctness of the above algorithm is clear by construction. We only need to show that it finishes in polynomial time.

Note that the above dynamic program runs in time $O(q|\mathcal{S}|)$. Now observe that q is bounded by n since there are at most n connected components. Furthermore, for each subset $\mathbb{C} \subseteq \{C_1, \dots, C_q\}$, we have that $\sum_{C \in \mathbb{C}} |C| \leq n$ (because the components in \mathbb{C} cannot contain more than n vertices). Thus, $|\mathcal{S}| \leq n + 1$ because each value $s \in \mathcal{S}$ corresponds to a subset of components $\mathbb{C} \subseteq \{C_1, \dots, C_q\}$ and each value $s \in \{0, \dots, n\}$ is only added once to \mathcal{S} . Hence, the algorithm runs in time $O(q|\mathcal{S}|) = O(n^2)$.

5.2.2 Computing Perfectly Balanced Assignments for Many Servers. We consider computing a perfectly balanced assignment respecting the connected components for ℓ servers.

Let C_1, \dots, C_q be the connected components which are assigned to the ℓ servers. To find a perfectly balanced assignment respecting the connected components, we need to find a partition of the set $\mathcal{S} = \{|C_1|, \dots, |C_q|\}$ into ℓ subsets $\mathcal{S}_1, \dots, \mathcal{S}_\ell$ such that for each subset \mathcal{S}_i we have that $\sum_{s \in \mathcal{S}_i} s = n/\ell$.

Unfortunately, the above problem is known to be NP-complete, see, e.g., the result about multi-processor scheduling in Garey and Johnson [14]. However, since we prove our results in the online model of computation, which allows unlimited computational power, the algorithm can solve this NP-complete problem. We note that this problem has also been studied in practice, see, e.g., Schreiber et al. [10] and references therein.

See Section 5.2.3 for how this problem can be solved approximately at the cost of a constant in the competitive ratio of the algorithm.

5.2.3 Computing Approximately Balanced Assignments for Many Servers. Previously we have seen that *perfectly* balanced assignments for ℓ servers cannot be computed in polynomial time unless $P = NP$ (Section 5.2.2). Thus, we now consider computing *approximately* balanced assignments for ℓ servers which is sufficient for our purpose: Let $\varepsilon' > 0$ be a constant. An assignment is $(1 + \varepsilon')$ -*approximately balanced* if each server has load at most $(1 + \varepsilon')n/\ell$. Using this definition, we obtain the following result.

PROPOSITION 32. *Let $\varepsilon > \varepsilon' > 0$ be constants and suppose each server has capacity $(1 + \varepsilon)n/\ell$. Then a $(1 + \varepsilon')$ -approximately balanced assignment for ℓ servers can be computed in polynomial time.*

Using the proposition (which we prove at the end of the subsection), we obtain a polynomial time algorithm with a slightly worse competitive ratio than that of Theorem 16.

THEOREM 33. *Given a system with ℓ servers each of capacity $(1 + \varepsilon)n/\ell$, for constant $\varepsilon \in (0, 1/2)$, then there exists an $O((\ell^2 \log n \log \ell)/\varepsilon^2)$ -competitive algorithm which runs in polynomial time.*

PROOF. First, observe that Algorithm 4 runs in polynomial time. Thus, the result of Proposition 22 also holds for polynomial time algorithms.

Second, consider a modification of Algorithm 1 where at each rebalancing step we compute a $(1 + \varepsilon')$ -approximately balanced assignment for $\varepsilon' = \varepsilon/2$. Such an approximately balanced assignment can be computed in polynomial time due to Proposition 32. Thus, the modified algorithm runs in polynomial time.

Observe that now all steps of the resulting algorithm can be computed in polynomial time. It is left to bound the competitive ratio of the modified algorithm.

We start by bounding the cost paid by the modified version of Algorithm 1. Note that each approximate rebalancing step incurs cost at most $O(\alpha n)$; recall that α denotes the cost for moving a vertex to a different server. Now we bound the number of approximate rebalancing steps. Recall from Lemma 4 that the number of vertex moves due to small-to-large steps is at most $O(n \log n)$. Now whenever a new approximately balanced assignment is computed, the small-to-large steps must have moved at least $\Omega((\varepsilon - \varepsilon')n/\ell)$ vertices to exceed the capacity of one of the servers. Thus, the total number of approximate rebalancing operations is bounded by $O((\ell \log n)/(\varepsilon - \varepsilon'))$ and, hence, the total cost of Algorithm 1 with approximate rebalancing steps is bounded by $O((\alpha n \ell \log n)/(\varepsilon - \varepsilon'))$.

Altogether, we obtain the following competitive ratio by following the steps from the proof of Theorem 16 (Section 4.5):

$$\frac{O(\alpha n \log n + (\alpha n \ell \log n)/(\varepsilon - \varepsilon'))}{\text{OPT}} = O\left(\frac{\alpha n \log n + (\alpha n \ell \log n)/(\varepsilon - \varepsilon')}{\alpha \varepsilon n / (\ell \log \ell)}\right) = O\left(\frac{\ell^2 \log n \log \ell}{\varepsilon^2}\right),$$

where in the last step we used that $\varepsilon - \varepsilon' = \varepsilon/2$. \square

To prove Proposition 32, we consider the makespan minimization problem in which there are k jobs with processing times p_1, \dots, p_k which must be assigned to ℓ identical machines. Given an assignment of the jobs to the machines, the maximum running time of any machine is called the *makespan*. The goal is to find an assignment of the jobs to the machines which minimizes the makespan.

The makespan minimization problem is known to be NP-hard but Hochbaum and Shmoys [15] presented a polynomial time approximation scheme (PTAS).

LEMMA 34 (HOCHBAUM AND SHMOYS [15]). *Let $\varepsilon' > 0$ be a constant. Then there exists an algorithm which computes a $(1 + \varepsilon')$ -approximate solution for the makespan minimization problem in polynomial time.*

Using the result from the lemma we can prove Proposition 32.

PROOF OF PROPOSITION 32. Suppose the system currently contains connected components C_1, \dots, C_k . We consider these connected components as the jobs of the makespan minimization problem with processing times $p_i = |C_i|$ for $i = 1, \dots, k$. The machines correspond to the ℓ servers.

Note that the optimal solution for the instance of the makespan minimization problem is n/ℓ : Since we have made the assumption that in the final assignment all servers have load exactly n/ℓ , there must exist a perfectly balanced assignment from the components C_i to the servers S_i . In other words, there exists an assignment of the jobs to the machines such that each machine has running time n/ℓ and, hence, the optimal makespan is n/ℓ .

By running the algorithm from Lemma 34, we obtain a $(1 + \varepsilon')$ -approximate solution for the makespan minimization problem. Since the optimal solution for this problem is n/ℓ , each machine has load at most $(1 + \varepsilon')n/\ell$ in the solution returned by the algorithm from Lemma 34. Assigning the components C_i to the servers in exactly the same way as the corresponding jobs are assigned to the corresponding machines, we obtain a $(1 + \varepsilon')$ -approximately balanced assignment in polynomial time. \square

6 LOWER BOUNDS

To study the optimality of our algorithms, we derive bounds on the competitive ratios which can be achieved by *any* deterministic online algorithm.

The following theorem provides a lower bound of $\Omega(1/\varepsilon + \log n)$. The lower bound has the following two main consequences: (1) If an algorithm is only allowed to use constant augmentation (i.e., servers of capacity $n/\ell + O(1)$), then the lower bound implies that any algorithm must have a competitive ratio of $\Omega(n)$.⁸ (2) The lower bound holds even in the setting in which there are only two servers. Thus, the algorithm from Section 3 for the two server setting is close to optimal (up to a $O(\min\{1/\varepsilon, \log n\})$ factor) and the generalized algorithm from Section 4 is optimal up to a $O(\ell \log \ell \min\{1/\varepsilon, \log n\})$ factor.

THEOREM 35. *Suppose there are two servers of capacity $(1 + \varepsilon)n/2$ for $\varepsilon \leq 0.98$. Then any deterministic online algorithm must have a competitive ratio of $\Omega(1/\varepsilon + \log n)$.*

To prove the theorem, we show in Section 6.1 that there exist input sequences such that *either* an algorithm always assigns vertices of the same connected component to the same server *or* it has prohibitively high cost. Using this fact, we prove our concrete lower bounds in Section 6.2.

6.1 Assigning Connected Components to Servers

In this subsection, we give an important reduction which will be useful to derive the lower bounds in the next subsection (Section 6.2). This reduction lets us assume that every competitive algorithm will always assign vertices of the same connected component to the same server.

More concretely, we show that every sequence of edges σ can be manipulated to a new edge sequence σ' such that: (1) σ reveals the same edges as σ' and (2) on input σ' , every algorithm *either* moves the vertices of the same connected components to the same server, *or* has prohibitively high cost and, hence, cannot be competitive.

We first prove the following technical lemma.

LEMMA 36. *Consider a sequence σ which reveals the edges $\emptyset \neq E^* \subseteq E$. Let C_1, \dots, C_q be the connected components induced by E^* .*

Then for each initial assignment there exists an input sequence σ' consisting only of edges in E^ such that either (1) at some point during the input sequence the algorithm assigns all vertices from each C_i to the same S_j or (2) the cost of the algorithm is at least $\Omega(\alpha n^3)$.*

PROOF. We will construct an input sequence σ' provided by the adversary such that either Property (1) or Property (2) must hold.

Consider an arbitrary initial assignment and pick the ground truth components V_i such that they do not coincide with the initial assignment of the vertices to the servers, i.e., $V_i \neq V_{\text{init}}(S_j)$ for all i, j . Let $E^* = \{e'_1, \dots, e'_t\}$ be the edges revealed by the adversary and suppose that E^* contains at least one edge (u, v) such that u and v are assigned to different servers in the initial assignment.

Now consider the input sequence $\sigma' = (e_1, \dots, e_r)$ with $r = \lceil \alpha n^3 t \rceil$ which consists of the edges (e'_1, \dots, e'_t) in E^* concatenated $\lceil \alpha n^3 \rceil$ times.

Suppose that while running the algorithm there always exists a C_i such that not all vertices from C_i are assigned to the same server S_j , i.e., Claim (1) does not apply. We show that then Claim (2) must apply.

Consider the state of the algorithm prior to a single subsequence containing the edges (e'_1, \dots, e'_t) . By assumption at least one edge e'_i must be between two vertices from different servers. Now the algorithm must either pay 1 for communication along this edge or it must move one of the edge's

⁸To obtain servers of capacity $n/\ell + O(1)$, we must set $\varepsilon = O(1)/n$.

endpoints at the cost of α to avoid paying for communication along this edge. Thus, the algorithm must pay at least $\Omega(1)$ for the subsequence (e'_1, \dots, e'_t) .

As there are $\lceil \alpha n^3 \rceil$ such subsequences, the algorithm must pay at least $\Omega(\alpha n^3)$ in total. \square

As we will see, the lemma essentially allows us to assume that every algorithm which obtains an edge between vertices on different clusters, must move their connected components to the same cluster. That is, given an input sequence σ , in our lower bound proof, we can employ Lemma 36 to obtain an input sequence σ' which does not reveal any additional edges and which forces every algorithm to have Property (1) or Property (2).

Now observe that if an algorithm has Property (2), since the cost of OPT are always bounded by $O(\alpha n)$ (OPT moves each vertex at most once), the algorithm cannot be competitive: the competitive ratio must be at least $\Omega(n^2)$, much higher than the competitive ratios derived in this paper. Hence, in the following we can assume that every algorithm with a competitive ratio better than $\Omega(n^2)$ must satisfy Property (1) of Lemma 36.

6.2 Lower Bound Proofs

In this subsection, we prove Theorem 35 by proving two different lower bounds: The first lower bound asserts a competitive ratio of $\Omega(1/\epsilon)$ and the second lower bounds asserts a competitive ratio of $\Omega(\log n)$.

In the lower bound constructions we heavily exploit that we provide hard instances against *deterministic* algorithms, i.e., we will rely on the fact that at each point in time the adversary knows exactly which assignment the online algorithm created.

Furthermore, we assume that after each edge which was provided by the adversary, the algorithm creates an assignment such that all vertices of the same connected component are assigned to the same server. This assumption is admissible by the discussion in Section 6.1.

We start by proving the lower bound of $\Omega(1/\epsilon)$.

LEMMA 37. *Consider the setting with two servers which both have capacity $(1 + \epsilon)n/2$ for $\epsilon > 0$.*

Then for each deterministic online algorithm ON there exists an input sequence σ such that the cost of ON is $\Omega(\alpha n)$ and the cost paid by OPT is $O(\alpha \epsilon n)$. Thus, the competitive ratio of every online algorithm is $\Omega(1/\epsilon)$.

PROOF. Choose an arbitrary initial assignment of n vertices to the ℓ servers. Let $K = \epsilon n/2$ denote the allowed augmentation of the servers. The initial assignment is as follows. In the left server, there are $q = n/(2(K + 1))$ connected components C_1, \dots, C_q of size $K + 1$. On the right server, we build one connected component of size $K + 1$ denoted C and one large connected component of size $n - K - 1$ denoted C' . First, the adversary provides all edges of these connected components at no cost to the algorithm.

Then the adversary inserts an edge from a vertex in C_1 to a vertex in C . Since C_1 has size $K + 1$ and the right server currently has $n/2$ vertices, the algorithm cannot move C_1 to the right server. For the same reason, the algorithm cannot move C to the left server either. Thus, the algorithm's only option to bring C_1 and C to the same server is to replace C with some C_i at the cost of $2\alpha(K + 1)$.

We will refer to the merged connected component of size $2(K + 1)$ as D . Note that D must be on the left server. Now let C_i be the connected component of size $K + 1$ on the right server. The adversary adds an edge from a vertex in D to a vertex in C_i . By the same reasoning as before, the algorithm must now pick some $C_j, j \neq i$, of size $K + 1$ from the left server and swap it with C_i . This costs another $2\alpha(K + 1)$.

The adversary continues the previous procedure until only a C_i of size $K + 1$ is left on the left server and then she connects C_i and C' . This gives the final partitioning of the vertices.

We observe that each vertex which is on the left server at the very end, has been on the right server exactly once during the execution of the algorithm. Thus, the costs paid by the algorithm must be $\Omega(\alpha n)$.

Note that OPT pays exactly $\alpha(K + 1)$ because it can determine beforehand which C_i must be moved to the right server and only move that connected component. Before, we have seen that any deterministic algorithm must pay at least $\Omega(\alpha n)$. Thus, the competitive ratio is $\Omega(n/K)$. \square

Next, we prove the $\Omega(\log n)$ lower bound for the competitive ratio of deterministic algorithms.

LEMMA 38. *Consider the setting with two servers which both have capacity $(1 + \varepsilon)n/2$ for $\varepsilon \leq 0.98$.*

Then for each deterministic online algorithm ON there exists an input sequence σ such that the cost paid by ON is $\Omega(\alpha n \log n)$ and the cost paid by OPT is $O(\alpha n)$. Thus, the competitive ratio of every deterministic online algorithm is $\Omega(\log n)$.

PROOF. Choose an arbitrary initial assignment of n vertices to the ℓ servers. Since we want to prove a lower bound, we can assume that n is a power of 2. Thus, suppose that $n = 2^a$ for $a \geq 1000$.

In our hard instance, we are creating a sequence of edge insertions which proceeds in $\Theta(\log n)$ rounds. When round i starts, all connected components have size 2^i induced by the previously provided edges, and when round i finishes, all connected components have size 2^{i+1} . We show that ON pays $\Omega(\alpha n)$ in each round. This implies the claimed cost of $\Omega(\alpha n \log n)$ for ON. The cost for OPT follows immediately from Lemma 3 which states that OPT never pays more than $O(n)$ when there are only two servers.

When ON starts and no edge was provided by the adversary, all connected components have size $1 = 2^0$, i.e., the connected components are isolated vertices.

Now suppose round $i = 0, \dots, \log n$ starts. By induction, all connected components have size 2^i . We now define a sequence of edge insertions for round i which forces ON to pay $\Omega(\alpha n)$ and after which all connected components have size 2^{i+1} .

Let z denote the current number of connected components of size 2^i . When round i starts, there are exactly $z = n/2^i = 2^{a-i}$ connected components of size 2^i each. Recall that each server has capacity $(1 + \varepsilon)n/2$. Thus, at most

$$y_i = (1 + \varepsilon)n/2^{i+1} \leq 1.98 \cdot 2^{a-i-1}$$

connected components of size 2^i can be assigned to each server.

Now suppose there exists an edge (u, v) such that C_u and C_v are of size 2^i and they are assigned to different servers; we call such an edge *expensive*. When the adversary inserts an expensive edge, ON must pay $\Omega(\alpha 2^i)$ for moving C_u or C_v to a different server.

The strategy of the adversary is to insert expensive edges as long as they exist. Once no expensive edges exist anymore, the adversary connects all remaining components of size 2^i arbitrarily until all components have size 2^{i+1} .

Note that expensive edges exist as long as $z > y_i$ (because when this inequality is satisfied, not all connected components of size 2^i can be assigned to the same server). Furthermore, observe that when the adversary inserts an expensive edge, z decreases by 2.

Now we prove a lower bound on the number of expensive edges p . By the previous arguments, p must be large enough such that:

$$z = 2^{a-i} - 2p \leq y_i.$$

Solving this inequality for p , we obtain

$$\begin{aligned} p &\geq 2^{a-i-1} - 1.98 \cdot 2^{a-i-2} \\ &= 2^{a-i-1}(1 - 0.99) \\ &= 0.01 \cdot 2^{a-i-1}. \end{aligned}$$

We conclude that that adversary can perform $\Omega(2^{a-i})$ expensive edge insertions. Since for each of these edge insertions, ON must pay $\Omega(\alpha 2^i)$, we obtain that the cost paid by ON in round i is

$$\Omega(\alpha \cdot 2^{a-i} \cdot 2^i) = \Omega(\alpha \cdot 2^{a-2}) = \Omega(\alpha n). \quad \square$$

7 SAMPLE APPLICATIONS: A DISTRIBUTED UNION FIND ALGORITHM AND ONLINE k -WAY PARTITIONING

In this section we provide two sample applications for our model and our algorithms. First, we show that our results can be used to solve a distributed union find problem and we give an example where a union find data structure is used in practice. Second, we show that our algorithms imply competitive algorithms for an online version of the k -way partitioning problem.

7.1 Distributed Union Find

Recall that in the *static* union find problem, there are n elements from a universe \mathcal{U} and initially there are n sets containing one element each. The data structure supports two operations: $\text{union}(u, v)$ and $\text{find}(u)$. Given two elements $u, v \in \mathcal{U}$, the operation $\text{union}(u, v)$ merges the sets containing u and v . The operation $\text{find}(u)$ returns the set containing u .

In the distributed setting we consider, elements are stored across ℓ servers. Each server has enough capacity to store $(1 + \epsilon)n/\ell$ elements and we have the natural constraint that elements from the same set must always be stored on the same server (in order to maintain locality for elements from the same set). We consider a setting in which all sets have size n/ℓ when the algorithm finishes.

Note that if the sets of $u, v \in \mathcal{U}$ are stored on different servers when the operation $\text{union}(u, v)$ is performed, one of the sets containing u or v must be moved to a different server. The goal of an algorithm is to minimize the moving cost caused by union-operations.

When analyzing the moving cost, we will compare with an optimal offline algorithm which knows in advance which union-operations will be performed. Thus, the optimal algorithm can move from the initial assignment to the final assignment at the minimum possible cost. For our analysis we will compute the competitive ratio between an online algorithm solving the above problem and the optimal offline algorithm (as also detailed in Section 2).

Using the algorithms from Sections 4 and 5.1, we obtain the following result.

THEOREM 39. *Consider a system with ℓ servers each of capacity $(1 + \epsilon)n/\ell$ for $\epsilon \in (0, 1/2)$. Then there exists a distributed $O((\ell \log n \log \ell)/\epsilon)$ -competitive algorithm for the distributed union find problem. Moreover, for $\ell = O(\sqrt{\epsilon n})$ servers, the algorithm's communication cost does not exceed its cost for moving vertices.*

PROOF. The theorem follows immediately from Theorem 29 by the following reduction from the model in Section 2. We identify vertices in the model from Section 2 with elements from the universe \mathcal{U} in the union find model. Furthermore, for each operation $\text{union}(u, v)$ we insert an edge (u, v) into the model from Section 2. Since all algorithms we considered always collocate vertices from the same connected component, they satisfy the constraint that elements from the same set must be assigned to the same server. Moreover, in our analysis we were able to focus on the number of vertex moves due to Lemma 1. In our proofs, we showed competitive bounds for the number

of vertex moves performed by the algorithm from Theorem 29 compared with an optimal offline algorithm. Thus, the same bounds as derived in Theorem 29 apply. \square

For $\ell = \Omega(\sqrt{\varepsilon n})$ servers and the exact number of messages sent by the algorithm, see Theorem 29. The guarantees from Theorem 29 carry over immediately.

An examples where distributed union find data structures are used in practice is search engines [13]. A search engine stores many different documents from the Web over multiple servers. Now union find data structures are used to collocate duplicate documents on the same server, i.e., when documents u and v are identified as duplicates the operation $\text{union}(u, v)$ is used to collocate these documents (and all previously identified duplicates) on the same server. Furthermore, union find data structures are used to find blocks in dense linear systems and in pattern recognition tasks (see Cybenko et al. [16] and references therein).

7.2 Online k -Way Partitioning

The model and algorithms we study in this paper can also be used to solve an online variant of the k -way partition problem [10]. In the static version of the k -way partition problem one is given a (multi-)set of integers \mathcal{S} and the task is to partition \mathcal{S} into k subsets $\mathcal{S}_1, \dots, \mathcal{S}_k$ such that the sum of all subsets is (approximately) equal.

Our model and our algorithms can be used to solve the following online version of this fundamental problem. Initially, \mathcal{S} contains n integers and all integers are 1. Each integer is assigned to one of ℓ bins and each bin has capacity $(1 + \varepsilon)n/\ell$. Now in an online sequence of operations, an adversary picks two integers from \mathcal{S} and these integers are added. For example, after adding integers $a, b \in \mathcal{S}$, \mathcal{S} becomes $\mathcal{S} = (\mathcal{S} \cup \{a + b\}) \setminus \{a, b\}$. During this sequence of operations an online algorithm must ensure that the load of all bins is always bounded by $(1 + \varepsilon)n/\ell$. We work under the assumption that after each operation there always exists an assignment from the integers in \mathcal{S} to the bins such that each bin has load exactly n/ℓ . We further assume that at the end of the sequence of operations there are ℓ integers and each integer is n/ℓ .

Note that when two integers $a, b \in \mathcal{S}$ from different bins are added, either a or b must be moved to a different bin. This might cause that bin to exceed its capacity.

We will analyze algorithms which have small moving cost. That is, the cost of an algorithm is the sum of the numbers it has moved. We consider the competitive analysis of online algorithms compared with an optimal offline algorithm which knows the sequence of additions in advance and which can move the numbers at optimal cost.

We then obtain the following result for the k -way partitioning problem.

THEOREM 40. *Consider a system with ℓ bins each of capacity $(1 + \varepsilon)n/\ell$ for $\varepsilon \in (0, 1/2)$. Then there exists a $O((\ell \log n \log \ell)/\varepsilon)$ -competitive algorithm for the k -way partition problem.*

PROOF. We can relate the online version of the k -way partition problem to the model we study by identifying integers and the sizes of connected components. Initially, we identify each $s \in \mathcal{S}$ with a single vertex. Note that this can be done since initially $s = 1$ and thus s and the size of its corresponding connected component are the same. After that, when two integers a and b are added, we take their corresponding connected components C_a and C_b and insert an edge between them. Note that the resulting integer $a + b$ corresponds to the connected component $C_a \cup C_b$ and their sizes agree, i.e., $a + b = |C_a \cup C_b|$. Now observe that summing the moving cost for integers is the same as counting the number of vertex reassignments for connected components. Thus, the result of the theorem follows from Theorem 16. \square

8 RELATED WORK

The design of more flexible networked systems that can adapt to their workloads has received much attention over the last years, with applications for traffic engineering [17, 18], load-balancing [19, 20], network slicing [21], server migration [22], switching [23, 24], or even adjusting the network topology [5]. The impact of distributed applications on the communication network is also well-documented in the literature [2, 3, 25–27]. Several empirical studies exploring the spatial and temporal locality in traffic patterns found evidence that these workloads are often *sparse and skewed* [5, 6, 28, 29], introducing optimization opportunities. E.g., studies of reconfigurable datacenter networks [5, 30] have shown that for certain workloads, a demand-aware datacenter network can achieve a performance similar to a demand-oblivious datacenter network at 25-40% lower cost [5, 30].

However, much less is known about the algorithmic challenges underlying such workload-adaptive networked systems, the focus of our paper. From an online algorithm perspective, our problem is related to reconfiguration problems such as online page (resp. file) migration [31, 32] as well as server migration [33] problems, k -server [34] problems, or online metrical task systems [35]. In contrast to these problems, in our model, requests do not appear somewhere in a graph or metric space but *between communication partners*. From this perspective, our problem can also be seen as a “distributed” version of online paging problems [9, 36–38] (and especially their variants *with bypassing* [39, 40]) where access costs can be avoided by moving items to a *cache*: in our model, access costs are avoided by collocating communication partners on the same *server* (a “distributed cache”).

The static version of our problem, how to partition a graph, is a most fundamental and well-explored problem in computer science [41], with many applications, e.g., in community detection [42]. The balanced graph partitioning problem is related to minimum bisection problems [43], and known to be hard even to approximate [44]. The best approximation today is due to Krauthgamer [45]. In contrast, we in this paper are interested in a dynamic version of the problem where the edges of the to-be-partitioned graph are revealed over time, in an online manner. Further, the offline problem of embedding workloads in a communication-efficient manner has been studied in the context of the minimum linear arrangement problem [46] and the virtual network embedding problem [47], however, without considering the option of migrations. In this regard, our paper features an interesting connection to the *itinerant list update* model [48], a kind of “dynamic” minimum linear arrangement problem which allows for reconfigurations and, notably, considers pair-wise requests. However, communication is limited to a linear line and so far, only non-trivial *offline* solutions are known.

One of the applications of the problem we study is a distributed union find data structure (see Section 7.1). Union find data structures have been initially proposed in the centralized setting and efficient algorithms were derived [11, 12]. Later, parallel versions of union find data structures were considered in a shared memory setting in which the goal was to derive wait-free algorithms [49]; also external memory algorithms were considered [50]. To the best of our knowledge studies of union find data structures in a distributed memory setting were only conducted experimentally, see (for example) [16, 51–53].

The second application we presented was as online k -way partitioning (Section 7.2). The k -way partitioning problem is known to be NP-hard as it constitutes a very simple scheduling problem [14]. The problem has also been researched in practice, see, e.g., [10, 54] and references therein. We are not aware of literature studying the online version of the problem which we have considered.

The paper most closely related to ours is by Avin et al. [8] who studied a more general version of the problem considered in our paper. In their model, request patterns can change arbitrarily over

time, and in particular, do not have to follow a partition and hence “cannot be learned”. Indeed, as we have shown in this paper, learning algorithms can perform significantly better: in [8], it was shown that for constant ℓ any deterministic online algorithm must have a competitive ratio of at least $\Omega(n)$ unless it can collocate *all* nodes on a single server, while we have presented an $O(\log n)$ -competitive online algorithm. Thus, our result is exponentially better than what can possibly be achieved in the model of [8].

9 CONCLUSION

Motivated by the increasing resource allocation flexibilities available in modern compute infrastructures, we initiated the study of online algorithms for adjusting the embedding of workloads according to the specific communication patterns, to reduce communication and moving costs. In particular, we presented algorithms and derived upper and lower bounds on their competitive ratio.

We believe that our work opens several interesting questions for future research. In particular, it remains to close the gap between the upper and lower bound of the competitive ratios derived in this paper. Furthermore, while in this paper we assumed that there are ℓ ground truth components of size n/ℓ , it will be interesting to study more general settings with smaller and larger components.

More generally, it will be interesting to consider algorithms which do not collocate all communication partners. Also, studying collocation in specific networks such as Clos networks, which are frequently encountered in datacenters, would be intriguing.

ACKNOWLEDGMENTS

We are grateful to our shepherd Rachit Agarwal as well as the anonymous reviewers whose insightful comments helped us improve the presentation of the paper.

The research leading to these results has received funding from the European Research Council under the European Community’s Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement No. 340506. Stefan Neumann gratefully acknowledges the financial support from the Doctoral Programme “Vienna Graduate School on Computational Optimization” which is funded by the Austrian Science Fund (FWF, project no. W1260-N35).

REFERENCES

- [1] M. Noormohammadpour and C. S. Raghavendra, “Datacenter traffic control: Understanding techniques and trade-offs,” *IEEE Communications Surveys & Tutorials*, 2017.
- [2] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B. Su, “Scaling distributed machine learning with the parameter server,” in *Proc. USENIX OSDI*, vol. 14, pp. 583–598, 2014.
- [3] J. C. Mogul and L. Popa, “What we talk about when we talk about cloud network performance,” *SIGCOMM Comput. Commun. Rev. (CCR)*, Sept. 2012.
- [4] C. Avin and S. Schmid, “Toward demand-aware networking: A theory for self-adjusting networks,” in *ACM SIGCOMM Computer Communication Review (CCR)*, 2018.
- [5] M. Ghobadi, R. Mahajan, A. Phanishayee, N. Devanur, J. Kulkarni, G. Ranade, P.-A. Blanche, H. Rastegarfar, M. Glick, and D. Kilper, “Projector: Agile reconfigurable data center interconnect,” in *Proc. ACM SIGCOMM*, (New York, NY, USA), pp. 216–229, ACM, 2016.
- [6] A. Roy, H. Zeng, J. Bagga, G. Porter, and A. C. Snoeren, “Inside the social network’s (datacenter) network,” in *Proc. ACM SIGCOMM*, Proc. ACM SIGCOMM, (New York, NY, USA), pp. 123–137, ACM, 2015.
- [7] T. Benson, A. Akella, and D. A. Maltz, “Network traffic characteristics of data centers in the wild,” in *Proc. ACM SIGCOMM Conference on Internet Measurement (IMC)*, IMC ’10, (New York, NY, USA), pp. 267–280, ACM, 2010.
- [8] C. Avin, A. Loukas, M. Pacut, and S. Schmid, “Online balanced repartitioning,” in *Proc. 30th International Symposium on Distributed Computing (DISC)*, 2016.
- [9] D. D. Sleator and R. E. Tarjan, “Amortized efficiency of list update and paging rules,” *Communications of the ACM*, vol. 28, no. 2, pp. 202–208, 1985.
- [10] E. L. Schreiber, R. E. Korf, and M. D. Moffitt, “Optimal multi-way number partitioning,” *J. ACM*, vol. 65, no. 4, pp. 24:1–24:61, 2018.

- [11] B. A. Galler and M. J. Fischer, "An improved equivalence algorithm," *Commun. ACM*, vol. 7, no. 5, pp. 301–303, 1964.
- [12] R. E. Tarjan and J. van Leeuwen, "Worst-case analysis of set union algorithms," *J. ACM*, vol. 31, no. 2, pp. 245–281, 1984.
- [13] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic clustering of the web," *Computer Networks*, vol. 29, no. 8–13, pp. 1157–1166, 1997.
- [14] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.
- [15] D. S. Hochbaum and D. B. Shmoys, "Using dual approximation algorithms for scheduling problems theoretical and practical results," *J. ACM*, vol. 34, no. 1, pp. 144–162, 1987.
- [16] G. Cybenko, T. G. Allen, and J. E. Polito, "Practical parallel union-find algorithms for transitive closure and clustering," *International Journal of Parallel Programming*, vol. 17, no. 5, pp. 403–423, 1988.
- [17] C.-Y. Hong, S. Kandula, R. Mahajan, M. Zhang, V. Gill, M. Nanduri, and R. Wattenhofer, "Achieving high utilization with software-driven wan," in *Proc. ACM SIGCOMM Computer Communication Review (CCR)*, vol. 43, pp. 15–26, 2013.
- [18] S. Jain, A. Kumar, S. Mandal, J. Ong, L. Poutievski, A. Singh, S. Venkata, J. Wanderer, J. Zhou, M. Zhu, *et al.*, "B4: Experience with a globally-deployed software defined wan," *Proc. ACM SIGCOMM Computer Communication Review (CCR)*, vol. 43, no. 4, pp. 3–14, 2013.
- [19] P. Patel, D. Bansal, L. Yuan, A. Murthy, A. Greenberg, D. A. Maltz, R. Kern, H. Kumar, M. Zikos, H. Wu, *et al.*, "Ananta: Cloud scale load balancing," in *Proc. ACM SIGCOMM Computer Communication Review (CCR)*, vol. 43, pp. 207–218, 2013.
- [20] D. E. Eisenbud, C. Yi, C. Contavalli, C. Smith, R. Kononov, E. Mann-Hielscher, A. Cilingiroglu, B. Cheyney, W. Shang, and J. D. Hosein, "Maglev: A fast and reliable software network load balancer," in *Proc. USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, pp. 523–535, 2016.
- [21] R. Sherwood, M. Chan, A. Covington, G. Gibb, M. Flajslik, N. Handigol, T.-Y. Huang, P. Kazemian, M. Kobayashi, J. Naous, *et al.*, "Carving research slices out of your production networks with openflow," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 1, pp. 129–130, 2010.
- [22] M. Bienkowski, A. Feldmann, J. Grassler, G. Schaffrath, and S. Schmid, "The wide-area virtual service migration problem: A competitive analysis approach," *IEEE/ACM Transactions on Networking (ToN)*, 2014.
- [23] P. Bosshart, D. Daly, G. Gibb, M. Izzard, N. McKeown, J. Rexford, C. Schlesinger, D. Talayco, A. Vahdat, G. Varghese, *et al.*, "P4: Programming protocol-independent packet processors," *Proc. ACM SIGCOMM Computer Communication Review (CCR)*, vol. 44, no. 3, pp. 87–95, 2014.
- [24] D. Firestone, "Smartnic: Accelerating azure's network with fpgas on ocs servers," <https://ocpusummit2016.sched.com/event/68u4/>, 2016.
- [25] C. Fuerst, S. Schmid, L. Suresh, and P. Costa, "Kraken: Online and elastic resource reservations for multi-tenant datacenters," in *Proc. 35th IEEE Conference on Computer Communications (INFOCOM)*, 2016.
- [26] A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannan, S. Boving, G. Desai, B. Felderman, P. Germano, *et al.*, "Jupiter rising: A decade of clos topologies and centralized control in google's datacenter network," *Proc. ACM SIGCOMM Computer Communication Review (CCR)*, vol. 45, no. 4, pp. 183–197, 2015.
- [27] Cisco, "Cisco global cloud index: Forecast and methodology, 2015-2020," *White Paper*, 2015.
- [28] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan, "Data center tcp (dctcp)," in *Proc. ACM SIGCOMM Computer Communication Review (CCR)*, vol. 40, pp. 63–74, 2010.
- [29] G. Judd, "Attaining the promise and avoiding the pitfalls of tcp in the datacenter.," in *Proc. USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, pp. 145–157, 2015.
- [30] N. Hamedazimi, Z. Qazi, H. Gupta, V. Sekar, S. R. Das, J. P. Longtin, H. Shah, and A. Tanwer, "Firefly: A reconfigurable wireless data center fabric using free-space optics," in *Proc. ACM SIGCOMM Computer Communication Review (CCR)*, vol. 44, pp. 319–330, 2014.
- [31] Y. Bartal, M. Charikar, and P. Indyk, "On page migration and other relaxed task systems," *Theoretical Computer Science*, vol. 268, no. 1, pp. 43–66, 2001. Also appeared in *Proc. of the 8th SODA*, pages 43–52, 1997.
- [32] D. L. Black and D. D. Sleator, "Competitive algorithms for replication and migration problems," 1989.
- [33] M. Bienkowski, A. Feldmann, J. Grassler, G. Schaffrath, and S. Schmid, "The wide-area virtual service migration problem: A competitive analysis approach," *IEEE/ACM Trans. Netw.*, vol. 22, pp. 165–178, Feb. 2014.
- [34] A. Fiat, Y. Rabani, and Y. Ravid, "Competitive k-server algorithms," *J. Comput. Syst. Sci.*, vol. 48, no. 3, pp. 410–428, 1994.
- [35] A. Borodin, N. Linial, and M. E. Saks, "An optimal on-line algorithm for metrical task system," *Journal of the ACM*, vol. 39, no. 4, pp. 745–763, 1992. Also appeared in *Proc. of the 19th STOC*, pages 373–382, 1987.
- [36] A. Fiat, R. M. Karp, M. Luby, L. A. McGeoch, D. D. Sleator, and N. E. Young, "Competitive paging algorithms," *Journal of Algorithms*, vol. 12, no. 4, pp. 685–699, 1991.
- [37] M. Mendel and S. S. Seiden, "Online companion caching," *Theoretical Computer Science*, vol. 324, no. 2–3, pp. 183–200, 2004.

- [38] N. E. Young, "On-line caching as cache size varies," in *Proc. of the 2nd ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pp. 241–250, 1991.
- [39] A. Adamaszek, A. Czumaj, M. Englert, and H. Räcke, "An $O(\log k)$ -competitive algorithm for generalized caching," in *Proc. 23rd SODA*, pp. 1681–1689, 2012.
- [40] L. Epstein, C. Imreh, A. Levin, and J. Nagy-György, "Online file caching with rejection penalties," *Algorithmica*, vol. 71, no. 2, pp. 279–306, 2015.
- [41] L. Vaquero, F. Cuadrado, D. Logothetis, and C. Martella, "Adaptive partitioning for large-scale dynamic graphs," in *Proc. 4th Annual Symposium on Cloud Computing (SOCC)*, pp. 35:1–35:2, 2013.
- [42] E. Abbe, "Community detection and stochastic block models: Recent developments," *Journal of Machine Learning Research*, vol. 18, no. 177, pp. 1–86, 2018.
- [43] U. Feige and R. Krauthgamer, "A polylogarithmic approximation of the minimum bisection," *SIAM Journal on Computing*, vol. 31, no. 4, pp. 1090–1118, 2002.
- [44] K. Andreev and H. Räcke, "Balanced graph partitioning," *Theory of Computing Systems*, vol. 39, no. 6, pp. 929–939, 2006.
- [45] R. Krauthgamer and U. Feige, "A polylogarithmic approximation of the minimum bisection," *SIAM Review*, vol. 48, no. 1, pp. 99–130, 2006.
- [46] S. Rao and A. W. Richa, "New approximation techniques for some ordering problems," in *SODA*, vol. 98, pp. 211–219, 1998.
- [47] M. Yu, Y. Yi, J. Rexford, and M. Chiang, "Rethinking virtual network embedding: substrate support for path splitting and migration," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 17–29, 2008.
- [48] N. Olver, K. Pruhs, K. Schewior, R. Sitters, and L. Stougie, "The itinerant list update problem," *Proc. 16th Workshop on Approximation and Online Algorithms (WAOA)*, 2018.
- [49] R. J. Anderson and H. Woll, "Wait-free parallel algorithms for the union-find problem," in *STOC*, pp. 370–380, 1991.
- [50] P. K. Agarwal, L. Arge, and K. Yi, "I/o-efficient batched union-find and its applications to terrain analysis," *ACM Trans. Algorithms*, vol. 7, no. 1, pp. 11:1–11:21, 2010.
- [51] F. Manne and M. M. A. Patwary, "A scalable parallel union-find algorithm for distributed memory computers," in *PPAM*, pp. 186–195, 2009.
- [52] M. M. A. Patwary, J. R. S. Blair, and F. Manne, "Experiments on union-find algorithms for the disjoint-set data structure," in *SEA*, pp. 411–423, 2010.
- [53] M. M. A. Patwary, P. Refsnes, and F. Manne, "Multi-core spanning forest algorithms using the disjoint-set data structure," in *IPDPS*, pp. 827–835, 2012.
- [54] R. E. Korf, "Multi-way number partitioning," in *IJCAI*, pp. 538–543, 2009.