

Near-Optimal Distributed Maximum Flow

Mohsen Ghaffari
Massachusetts Institute of Technology
Cambridge MA, USA
ghaffari@csail.mit.edu

Andreas Karrenbauer
MPI for Informatics
Saarbrücken, Germany
karrenba@mpi-inf.mpg.de

Fabian Kuhn
University of Freiburg
Freiburg, Germany
kuhn@cs.uni-freiburg.de

Christoph Lenzen
MPI for Informatics
Saarbrücken, Germany
clenzen@mpi-inf.mpg.de

Boaz Patt-Shamir
Tel Aviv University
Tel Aviv, Israel
boazps@post.tau.ac.il

Abstract

We present a near-optimal distributed algorithm for $(1 + o(1))$ -approximation of single-commodity maximum flow in undirected weighted networks that runs in $(D + \sqrt{n}) \cdot n^{o(1)}$ communication rounds in the CONGEST model. Here, n and D denote the number of nodes and the network diameter, respectively. This is the first improvement over the trivial bound of $O(n^2)$, and it nearly matches the $\tilde{\Omega}(D + \sqrt{n})$ round complexity lower bound.

The development of the algorithm contains two results of independent interest:

- (i) A $(D + \sqrt{n}) \cdot n^{o(1)}$ -round distributed construction of a spanning tree of average stretch $n^{o(1)}$.
- (ii) A $(D + \sqrt{n}) \cdot n^{o(1)}$ -round distributed construction of an $n^{o(1)}$ -congestion approximator consisting of the cuts induced by $O(\log n)$ virtual trees. The distributed representation of the cut approximator allows for evaluation in $(D + \sqrt{n}) \cdot n^{o(1)}$ rounds.

All our algorithms make use of randomization and succeed with high probability.

1 Introduction

Computing a maximum flow is a fundamental task in network optimization. While the problem has a decades-old history rich with developments and improvements in the sequential setting, little is known in the distributed setting. In fact, prior to this work, the best known distributed time complexity in the standard CONGEST model remained at the trivial bound of $O(m)$, which is the time needed to collect the entire topology and solve the problem locally (see more in Section 1.2). For undirected networks, this paper improves this unsatisfying state to near-optimality:

Theorem 1.1. *On undirected weighted graphs, a $(1 + \varepsilon)$ -approximation of a maximum s - t flow can be computed in $(D + \sqrt{n}) \cdot n^{o(1)} \varepsilon^{-3}$ rounds of the CONGEST model with high probability.*

This round complexity almost matches the $\tilde{\Omega}(D + \sqrt{n})$ lower bound of Das Sarma et al. [14], which holds for any non-trivial approximation. Before we proceed, let us formalize the model and the problem.

1.1 Model and Problem

Model. We use the standard CONGEST model of synchronous computation [24]. We are given a simple, connected, weighted graph $G = (V, E, \text{cap})$, where $\text{cap} : E \rightarrow \mathbb{N}$, $\text{cap}(e) \in \text{poly } n$, are the edge capacities.

ities.¹ By D , we denote the (hop) diameter of G . Each of the $n := |V|$ nodes hosts a processor with a unique identifier of $O(\log n)$ bits, and over each of the $m := |E|$ edges $O(\log n)$ bits can be sent in each synchronous round of communication; we assume that nodes have access to infinite strings of independent unbiased random bits. We say that an event occurs *with high probability* (w.h.p.), if it happens with probability $1 - n^{-c}$ for any desired constant $c > 0$ specified upfront.² Initially, each node only knows its identifier, its incident edges, and their capacities.

Problem. We fix an arbitrary orientation of the edges. In the following, we write $(u, v) \in E$ if $\{u, v\} \in E$ is directed from u to v . An instance of the (single-commodity) *max flow* problem is given by, in addition to specifying G , designating a source $s \in V$ and a sink $t \in V$. A (feasible) *flow* is a vector $\mathbf{f} \in \mathbb{R}^E$ satisfying:

1. capacity constraints (edges): $\forall e \in E : |f_e| \leq \text{cap}(e)$;
2. preservation constraints (nodes): $\forall u \in V \setminus \{s, t\} : \sum_{(u,v) \in E} f_e - \sum_{(v,u) \in E} f_e = 0$; and
3. $\sum_{(s,u) \in E} f_e - \sum_{(u,s) \in E} f_e = -\sum_{(t,u) \in E} f_e + \sum_{(u,t) \in E} f_e = F \in \mathbb{R}$.

Here, F is the *value* of \mathbf{f} . A *max flow* is a flow of maximum value. For $\varepsilon > 0$, a $(1 + \varepsilon)$ -*approximate max flow* is a flow whose value is by at most a factor $1 + \varepsilon$ smaller than that of a max flow. In this work, we focus on solving the problem of finding a $(1 + \varepsilon)$ -approximate max flow in the above model, where it suffices that each node u learns f_e for its incident edges $\{u, v\} \in E$.

1.2 Related Work

Network flow, being one of the canonical and most useful optimization problems, has been the target of innumerable research efforts since the 1930s [27] (see, e.g., the classic book [2] and the recent survey [15]). For the general, directed case, the fastest known sequential algorithm is by Goldberg and Rao and it solves the max flow problem in time $\tilde{O}(m \cdot \min\{m^{1/2}, n^{1/2}\})$. Particularly relevant from the point of view of the present paper are recent efforts to obtain fast algorithms to compute (approximate) max flow solutions in the undirected case. Using the graph sparsification technique of Benczúr and Karger [11], any graph can be partitioned into $k = \tilde{O}(m\varepsilon^2/n)$ sparse graphs with $\tilde{O}(n/\varepsilon^2)$ edges such that the max flow problem can be approximately solved by combining max flow solutions for each of these sparse graphs. Using the algorithm of Goldberg and Rao, this results in an algorithm with running time $\tilde{O}(mn^{1/2})$. In [13], Christiano et al. improved this to $\tilde{O}(mn^{1/3})$ by applying the almost linear-time Laplacian solver of Spielman and Teng [30] to iteratively minimize a softmax approximation of the edge congestions. Finally, (independently) in [16] and in [29], two algorithms appeared which allow to compute a $(1 + \varepsilon)$ -approximation to an undirected max flow problem in time almost linear in m .

However, to the dismay of many, and despite the fact that the word “network” even appears in the problem’s name, only little progress was made over the years from the standpoint of distributed algorithms. For example, Goldberg and Tarjan’s push-relabel algorithm, which is very local and simple to implement in the CONGEST model, requires $\Omega(n^2)$ rounds to converge, where n is the number of nodes. This is very disappointing, because in the CONGEST model, any problem whose input and output can be encoded with $O(\log n)$ bits per edge, can be trivially solved in $O(m)$ rounds, where m is the number of edges, by collecting all input at a single node, solving it there, and distributing the results back.

Early attempts focused, as customary in those days, on reducing the number of messages in asynchronous executions. For example, Segall [28] gives an $O(nm^2)$ -messages, $O(n^2m)$ -time algorithm for exact max flow, and Gafni and Marberg [21] give an algorithm whose message and time complexities are $O(n^2m^{1/2})$. Awerbuch has attacked the problem repeatedly with the following results. In an early work [6]

¹As merely an approximate flow is required, we can reduce the general case to this setting in $\tilde{O}((\sqrt{n} + D) \log C)$ rounds, where C is an upper bound on the ratio between the largest and smallest capacity.

²Taking the union bound over polynomially many events does not affect this property. We will use this fact frequently and implicitly throughout the paper.

he adapts Dinic’s centralized algorithm using a synchronizer, giving rise to an algorithm whose time and message complexities are $O(n^3)$. With Leighton, in [9] they give an algorithm for solving multicommodity flow approximately in $O(\ell m \log m)$ rounds, where $\ell < n$ is the length of the longest flow path. Later he considers the model where each flow path (variable) has an “agent” which can find the congestion of all links on its path in *constant time*. In this model, he shows with Khandekar [7] how to approximate any positive LP (max flow with given routes included) to within $(1 - \epsilon)$ in time polynomial in $\log(mnA_{\max}/\epsilon)$ (here n is the number of variables, which is at least the number of paths considered). The same model is used with Khandekar and Rao in [8], where they show how to approximate multicommodity flow to within $(1 - \epsilon)$ in $O(\ell \log n)$ rounds. Using a straightforward implementation of this algorithm in the CONGEST model results in an $\tilde{O}(n^2)$ -time algorithm.

Thus, up to the current paper, there was no distributed implementation of a max-flow algorithm which always requires subquadratic number of rounds. Even an $O(n)$ -time algorithm would have been considered a significant improvement, even for the 0/1 capacity case.

1.3 Organization of this Article

Our result builds heavily on a few major breakthroughs in the understanding of max flow in the centralized setting, most notably [29], as well as a few other contributions. We first give an overview of the key concepts in Sections 2–4. We carefully revisit Sherman’s approach [29] and the main building blocks he relies on in Section 2. This sets the stage for shedding light on the challenges that must be overcome for its distributed implementation and presentation of our results in Section 3. There, we also provide a top-level view of the components of the algorithm, alongside pointers to the detailed proofs in Sections 5–9 showing that we can implement each of them by efficient distributed algorithms. In Section 4, we outline the distributed construction of an $n^{o(1)}$ -congestion approximator, which is our key technical contribution; the role of a congestion approximator is to *estimate* the congestion induced by optimally routing an arbitrary demand vector very quickly, which lies at the heart of the algorithm. All the details of our distributed algorithm and all the proofs appear in Sections 5–9.

2 Overview of the Centralized Framework

Sherman’s approach [29] is based on *gradient descent* (see, e.g., [22]) for *congestion minimization* with a clever dualization of the flow conservation constraints. The flow problem is re-formulated as a *demand vector* $\mathbf{b} \in \mathbb{R}^n$ such that $\sum_{i \in V} b_i = 0$. In the case of the s - t flow problem, we have a positive b_s and negative b_t with the same absolute value and the demand is zero everywhere else. The objective is to find a flow \mathbf{f}^* that meets the given demand vector, i.e., the total excess flow in node i is equal to b_i , and minimizes the maximum *edge congestion*, which is the ratio of the flow over an edge to its capacity. Formally:

$$\text{minimize } \|C^{-1} \mathbf{f}\|_{\infty} \text{ subject to } B\mathbf{f} = \mathbf{b}, \quad (1)$$

where $C = (C_{ee'})_{e,e' \in E}$ is an $m \times m$ diagonal matrix with

$$C_{ee'} = \begin{cases} \text{cap}(e) & \text{if } e = e' \\ 0 & \text{else,} \end{cases}$$

and $B = (B_{ve})_{v \in V, e \in E}$ is an $n \times m$ matrix with

$$B_{ve} = \begin{cases} 1 & \text{if } e = (u, v) \text{ for some } u \in V \\ -1 & \text{if } e = (v, u) \text{ for some } u \in V \\ 0 & \text{else.} \end{cases}$$

Note that given a general (i.e., unconstrained) flow vector $\mathbf{f} \in \mathbb{R}^m$, $(B\mathbf{f})_v$ is exactly the excess flow at node v . Hence, by the max-flow min-cut theorem, if we can solve problem (1), a simple binary search will find an approximate max flow.

Instead of directly solving this constrained system, Sherman allows for general flows and adds a penalty term for any violation of flow constraints, i.e.,

$$\text{minimize } \|C^{-1}\mathbf{f}\|_\infty + 2\alpha \|R(\mathbf{b} - B\mathbf{f})\|_\infty ,$$

where $\alpha \geq 1$ and the matrix R are chosen so that the optimum of this unconstrained optimization problem does not violate the flow constraints. As we are interested in an approximate max flow, we can compute an approximate solution and argue that the violation of the flow constraints will be small, too. Then one simply re-routes the remaining flow in a trivial manner, e.g. on a spanning tree, to obtain a near-optimal solution. Finally, to ensure that the objective function is differentiable (i.e., a gradient descent is actually possible), $\|\cdot\|_\infty$ is replaced by the so-called soft-max.

The Congestion Approximator R . The *congestion* of an edge e (for a given flow \mathbf{f}) is defined as the ratio $|f_e|/\text{cap}(e)$. When referring to the congestion of a cut in a given flow, we mean the ratio between the net flow crossing the cut to the total capacity of the cut. Suppose for a moment that $\alpha = 1$ and R contains one row for each cut of the graph, chosen such that each entry of the vector $RB\mathbf{f}$ equals the congestion of the corresponding cut. In particular, R would correctly reproduce the congestion of min cuts (which give rise to maximal congestion). Moreover, the vector $R\mathbf{b}$ describes the inevitable congestion of the cuts for any feasible flow. Thus, the components of $R(\mathbf{b} - B\mathbf{f})$ are the residual congestions to be dealt with to make \mathbf{f} feasible (neglecting possible cancellations). The max-flow min-cut theorem and the factor of 2 in the second term of the objective function imply that it always improves the value of the objective function to route the demands arising from a violation of flow constraints optimally. Moreover, the gradient descent concentrates on the most congested edges and those that are contained in cuts with the top residual congestion. In particular, flow is pushed over the edges into the cut with the highest residual congestion to satisfy its demand until other cuts become more important in the second part of the objective. The first part of the objective impedes flow on edges the more they are congested (on an absolute scale and relative to others). Thus, approximately minimizing the objective function is equivalent to simultaneously approximating the minimum congestion and having small violation of flow constraints; solving up to polynomially small error and naively resolving the remaining violations then yields sufficiently accurate results.

Unfortunately, trying to make R capture congestion *exactly* is far too inefficient. Instead, one uses an α -congestion approximator, that is a matrix R such that for any demand vector \mathbf{b} , it holds that

$$\|R\mathbf{b}\|_\infty \leq \text{opt}(\mathbf{b}) \leq \alpha \|R\mathbf{b}\|_\infty ,$$

where $\text{opt}(\mathbf{b})$ is the maximum congestion caused on any cut by optimally routing \mathbf{b} . Since the second term in the objective function is scaled up by factor α , we are still guaranteed that optimally routing any excess demands improves the objective function. However, this implies that the second term of the objective function may dominate its gradient and thus emphasis is shifted rather to feasibility than optimality. Sherman proves that this slows down the gradient descent by at most a factor of α^2 , i.e., if $\alpha \in n^{o(1)}$, so is the number of iterations of the gradient descent algorithm that need to be performed.

Congestion Approximators: Racke’s Construction. For any spanning tree T of G , deleting an edge partitions the nodes into two connected components and thus induces an (edge) cut of G . Note that on T , this cut contains only the single deleted edge, and in terms of congestion any cut of T is dominated by such an edge-induced cut: For any cut, the maximum congestion of an edge is at least the average congestion of the cut, and in T , there is a cut containing only this edge.

These basic properties motivate the question of how well the cut structure of an arbitrary graph can be approximated by trees. Intuitively, the goal is to find a tree T (not necessarily a subgraph) spanning all nodes

with edge weights such that routing *any* demand vector in G and in T results in roughly the same maximal congestion. Because routing flows on trees is trivial, such a tree T would give rise to an efficient congestion approximator R : R would consist of one row for each cut induced by an edge (u, v) of T with capacity C , where the matrix entry corresponding to node w is $1/C$ if w is on u 's "side" of the cut and 0 otherwise; multiplying a demand vector with the row then yields the flow that needs to pass through (u, v) divided by the capacity of the cut.

In a surprising result [26], Räcke showed that, using multiplicative weight updates (see e.g. [5, 25, 31]) one can construct a distribution of $\tilde{O}(m)$ trees so that (i) in each tree of the distribution, each cut has at least the same capacity as in G and (ii) given any cut of G of total capacity C , sampling from the distribution results in a tree T where this cut has *expected* capacity $O(\alpha C)$; here α is the approximation ratio of a *low average stretch spanning tree* algorithm Räcke's construction uses as subroutine. Note that this bound on the expectation implies that for any cut of capacity C , there must be a tree in the distribution for which the cut has capacity $O(\alpha C)$. Hence, the cuts given by *all* trees in the distribution give rise to an $O(\alpha)$ -congestion approximator R with $\tilde{O}(mn)$ rows.

Low Average Stretch Spanning Trees. In order to perform Räcke's construction, one requires an efficient algorithm for computing low average stretch spanning trees. More precisely, given a graph $G = (V, E, \ell)$ with polynomially bounded lengths $\ell : E \rightarrow \mathbb{N}$, the goal is to construct a spanning tree T of G so that

$$\sum_{\{u,v\} \in E} d_T(u,v) \leq \alpha \sum_{\{u,v\} \in E} \ell(\{u,v\}),$$

where $d_T(u, v)$ is the sum of the lengths of the unique path from u to v in T and α is the *stretch* factor.

Sherman's algorithm builds on a sophisticated low average stretch spanning tree algorithm that achieves $\alpha \in O(\log n \log^2 \log n)$ within $\tilde{O}(m)$ centralized steps [1]. We use a simpler approach providing $\alpha \in 2^{O(\sqrt{\log n \log \log n})}$ [3] that has been shown to parallelize well, i.e., has an efficient implementation in the PRAM model [12].

Congestion Approximators: Madry's Construction. Räcke's construction has the drawback that one needs to sequentially compute a linear number of trees, which is prohibitively expensive from our point of view. Madry generalized Räcke's approach to a construction that results in a distribution over $\tilde{O}(m/j)$ so-called *j-trees* [19], where j is a parameter. A *j-tree* consists of a *forest* of j connected components (trees) and a *core graph*, which is an arbitrary connected graph with j nodes: one from each tree (see Figure 1).

The properties of the distribution are the same as for Räcke's: sampling from the distribution preserves cut capacities up to an expected $O(\alpha)$ -factor, where α is the stretch of the utilized spanning tree algorithm. Likewise, using all (dominant) cuts of all *j-trees* in the distribution to construct R yields an $O(\alpha)$ -congestion approximator. Note that any cut in a *j-tree* is *dominated* by either a cut induced by an edge of the forest, or by a cut of the core, in the following sense: Consider any demand vector and any "mixed" cut. If there is an edge in the forest crossing the cut that has at least the same congestion as the whole cut, then the cut induced by the forest edge dominates the mixed cut. Otherwise, we can remove all forest edges from the mixed cut without reducing its congestion. As routing demands in the forest part of the graph is trivial, Madry's construction can be seen as an efficient reduction of the problem size.

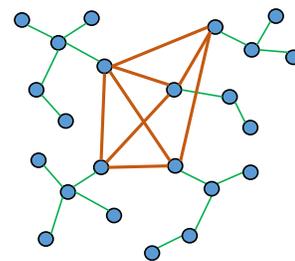


Figure 1: A j -tree for $j = 5$. The core links are depicted in brown.

Congestion Approximators: Combining Cut Sparsifiers with Madry's Construction. Using *j-trees*, Sherman derives a suitable congestion approximator, i.e., one with $\alpha \in n^{o(1)}$ that can be constructed and

evaluated in $\tilde{O}(m + n^{1+o(1)})$ rounds, as follows. First, a *cut sparsifier* is applied to G . A $(1 + \varepsilon)$ -sparsifier computes a subgraph of G with modified edge weights so that the capacities of all cuts are preserved up to factor $1 + \varepsilon$. It is known how to compute a $(1 + o(1))$ -sparsifier with $\tilde{O}(n)$ edges in $\tilde{O}(m)$ steps using randomization [11]. As the goal is merely to compute a congestion approximator with $\alpha \in n^{o(1)}$, the multiplicative $1 + o(1)$ approximation error is negligible. Hence, this essentially breaks the problem of computing a congestion approximator down to the same problem on sparse graphs.

Next, Sherman applies Madry's construction with $j = n/\beta$, where $\beta = 2^{\sqrt{\log n}}$. This yields a distribution of $\tilde{O}(\beta)$ many n/β -trees. The issue is now that the cores are arbitrary graphs, implying that it may be difficult to evaluate congestion for cuts in the cores. However, the number of nodes in the core is $n' = n/\beta$. Thus, recursion does the trick: apply the cut sparsifier to the core, use Madry's construction on the resulting graph (with $j' = n'/\beta = n/\beta^2$), rinse and repeat. In total, there are $\log_\beta n = \sqrt{\log n}$ levels of recursion until the core becomes trivial, i.e., we arrive at a tree. For each level of recursion, the approximation ratio deteriorates by a multiplicative $\alpha \in \text{polylog } n$, where α is the stretch factor of the low-stretch spanning tree algorithm, and a multiplicative $1 + o(1)$, for applying the cut sparsifier. This yields an α' -congestion approximator with

$$\alpha' \in ((1 + o(1))\alpha)^{\sqrt{\log n}} \subset 2^{O(\sqrt{\log n} \log \log n)} \subset n^{o(1)}.$$

While the total number of constructed trees is still $\tilde{O}(\beta^{\log_\beta n}) = \tilde{O}(n)$, the number of nodes in a graph (i.e., a core from the previous level) on the i^{th} level of recursion is only n/β^{i-1} . The cut sparsifier ensures that the number of edges in this graph is reduced to $\tilde{O}(n/\beta^{i-1})$ before recursing. Since the number of edges in the core is (trivially) bounded by the number of edges of the graph in Madry's construction, the total number of sequential computation steps for computing the distribution is thus bounded by

$$\tilde{O}(m) + \sum_{i=1}^{\log_\beta n} \tilde{O}(\beta^i \cdot n/\beta^{i-1}) \subset \tilde{O}(m + n^{1+o(1)}).$$

Step Complexity of the Flow Algorithm. The above recursive structure can also be exploited to *evaluate* the α' -congestion approximator Sherman uses in $n^{1+o(1)}$ steps. As mentioned earlier, the cuts of a j -tree are dominated by those induced by edges of the forest and those which are crossed by core edges only (cf. Figure 1). In the forest component, routing demands is unique, takes linear time in the number of nodes (simply start at the leaves), and results in a modified demand vector at the core on which is recursed.

Sherman proves that his algorithm obtains a $(1 + \varepsilon)$ -approximate flow in $O(\varepsilon^{-3}\alpha^2 \log^2 n)$ gradient descent steps, provided R is an α -congestion approximator.³ It is straightforward to see (cf. Section 9.1) that each of these steps requires $O(m)$ computational steps besides doing two matrix-vector multiplications with R and R^\top , respectively. Using the above observation and plugging in the time to construct the (implicit) representation of R , one arrives at a total step complexity of $\tilde{O}(mn^{o(1)})$.

3 Distributed Algorithm: Contribution and Key Ideas

The Distributed Toolchain. For a distributed implementation of Sherman's approach, many subproblems need to be solved (sufficiently fast) in the CONGEST model. We summarize them in the following list, where stars indicate that these components are readily available from prior work.

- * Decomposing trees into $O(\sqrt{n})$ components of strong diameter $O(\sqrt{n})$, within $\tilde{O}(\sqrt{n} + D)$ rounds. This can, e.g., be done by techniques pioneered by Kutten and Peleg for the purpose of minimum-weight spanning tree construction [18].

³Sherman points out that using Nesterov's accelerated gradient descent method [23], this can be improved to $O(\varepsilon^{-2}\alpha \log^2 n)$. For both his and our results, this difference is insubstantial, as $\alpha \in n^{o(1)} \Leftrightarrow \alpha^2 \in n^{o(1)}$.

- * Constructing cut sparsifiers. Koutis [17] provides a solution that completes in polylog n rounds of the CONGEST model. In Section 6, we give a simulation result for use in the recursive construction.
- 1. Constructing low average stretch spanning trees on multigraphs (Section 7).
- 2. Applying Madry’s construction in the CONGEST model, even when recursing in the context of Sherman’s framework (Section 8).
- 3. Sampling from the recursively constructed distribution (Section 8).
- 4. Avoiding the use of the entire distribution for constructing the congestion approximator (see below).
- 5. Performing a gradient descent step. This involves, e.g., matrix-vector multiplications with R , R^\top and C^{-1} , evaluation of the soft-max, etc. (Section 9).

We next present some additional description for the items 1 to 5 in the above list.

1. Low Average Stretch Spanning Trees. In Section 7, we prove the following theorem.

Theorem 3.1. *Suppose H is a multigraph obtained from G by assigning arbitrary edge lengths in $[2^{n^{o(1)}}]$ to the edges of G (known to incident nodes) and performing an arbitrary sequence of contractions. Then we can compute a spanning tree of H of expected stretch $2^{O(\sqrt{\log n \log \log n})}$ within $(\sqrt{n} + D)n^{o(1)}$ rounds.*

To obtain this theorem, we translate a PRAM algorithm by Blelloch et al. [12] to the CONGEST model. The main issue when transitioning from the PRAM to the CONGEST model is that in the PRAM model, information about distant parts of the graph may be readily accessed. In the CONGEST model, we handle this by pipelining long-distance communication over a global breadth-first-search (BFS) tree of G ; communication over $O(\sqrt{n})$ hops is handled using the edges that have already been selected for inclusion into the spanning tree and spanning trees of the contracted regions of G .

2. Implementing Madry’s Scheme. This is technically the most challenging part. Also here, we have to overcome the difficulty of potentially needing to communicate a large amount of information over many hops; doing this naively results in too much contention and thus slow algorithms. We approach this by modifying Madry’s construction so that:

- Instead of “aggregating” edges so that the core becomes a graph, we admit a multigraph as core.
- We do not explicitly construct the core. Instead, we simulate both the sparsifier and the low average stretch spanning tree algorithm using the abstraction of *cluster graphs* (see Section 5).
- In doing so, we maintain that every core edge is also a graph edge. This enables to handle all communication over this edge by using the corresponding graph edge.
- In this context, clusters are the forest components rooted at core nodes. We will maintain that forest components have depth $\tilde{O}(\sqrt{n})$. While this is not strictly necessary, it simplifies the description of the corresponding distributed algorithms, as the communication within each cluster can then be performed via its (previously constructed) spanning tree.
- The cluster hierarchy that is established during the construction allows for a straightforward recursive evaluation of the corresponding congestion approximator.

Section 8 gives the details of the construction.

3. Sampling from the Distribution. This is now straightforward, because for each sample, on each level of the recursion we need to construct only $n^{o(1)}$ different j -trees for some j . This is also discussed in Section 8, in which the formal version of the following theorem is proved.

Theorem 3.2 (Informal). *Within $\tilde{O}((\sqrt{n} + D)\beta)$ rounds of the CONGEST model, we can sample a virtual tree from the distribution used in Sherman’s framework, where $\tilde{O}(\beta)$ is the number of j -trees in the distribution constructed when recursing on a core. The distributed representation allows to evaluate the dominant cuts of the tree when using it in a congestion approximator within $\tilde{O}(\sqrt{n} + D)$ rounds.*

4. Avoiding the use of the entire distribution for constructing the congestion approximator. While Sherman can afford to use all trees in the (recursively constructed) distribution, the above theorem is not strong enough to allow for fast evaluation of all $\tilde{\Theta}(n)$ trees. As Madry points out [19], it suffices to sample and use $O(\log n)$ j -trees from the distribution he constructs to speed up any β -approximation algorithm for an “undirected cut-based minimization problem”, at the expense of an increased approximation ratio of $2\alpha\beta$, where α is the approximation ratio of the congestion approximator corresponding to the distribution of j -trees. The reasoning is as follows:

- The number of cuts that need to be considered for such a problem is polynomially bounded.
- The expected approximation ratio for any fixed cut when sampling from the distribution is α . By Markov’s bound, with probability at least $1/2$ it is at most 2α .
- For $O(\log n)$ samples, the union bound shows that w.h.p. *all* relevant cuts are 2α -approximated.
- Applying a β -approximation algorithm relying on the samples only, which can be evaluated much faster, results in a $2\alpha\beta$ -approximation w.h.p.

Recall that the problem of approximating a max flow was translated to minimizing congestion for demands F and $-F$ at s and t and performing binary search over F . The max-flow min-cut theorem implies the respective congestion to be the function of a single cut, which can be used to verify that the problem falls under Madry’s definition.

Unfortunately, applying the sampling strategy as indicated by Madry is infeasible in Sherman’s framework. As the goal is a $(1 + \varepsilon)$ -approximation, applying it to the above problem directly will yield a too inaccurate approximation. Alternatively, we can apply it in the construction of a congestion approximator. However, a congestion approximator must return a good approximation for *any* demand vector. There are exponentially many such vectors even if we restrict $\mathbf{b} \in \{-1, 0, 1\}^V$, and we are not aware of any result showing that the number of min-cuts corresponding to the respective optimal flows is polynomially bounded.

We resolve this issue with the following simple, but essential insight, at the expense of squaring the approximation ratio of the resulting congestion approximator.

Lemma 3.3. *Suppose we are given a distribution of poly n trees so that given any cut of G of capacity C , sampling from the distribution results in a tree whose corresponding cut has at least capacity C and at most capacity αC in expectation. Then sampling $O(\log n)$ such trees and constructing a congestion approximator from their single-edge induced cuts results in a $2\alpha^2$ -congestion approximator of G w.h.p.*

Proof. Recall that a cut approximator estimates the maximum congestion when optimally routing an arbitrary demand. Consider any demand vector and denote by C the capacity of the corresponding cut that is most congested when routing the demand. As sampling from the distribution yields approximation factor α in expectation, there must be *some* tree T in the distribution whose corresponding cut has capacity at most αC . However, this means that when routing the demand via T , there is some edge in T that experiences at least $1/\alpha$ times the maximum congestion when routing the demand optimally in G . As the capacity of the edge is at least that of the corresponding cut in G , it follows that the corresponding cut of G has congestion at least $1/\alpha$ of that of the min-cut when routing the demand.

As there are poly n trees, each of which has $n - 1$ edges, this shows that for any demand vector there is one of polynomially many cuts of G that experience at least $1/\alpha$ times the maximum congestion when optimally routing the demand vector. By Markov’s bound and the union bound, w.h.p. the congestion on each of these cuts will be approximated up to another factor of 2α when using $O(\log n)$ samples. \square

5. Performing a gradient descent step. Most of the high-level operations required for executing a gradient descent algorithm are straightforward to implement using direct communication between neighbors or broadcast and convergecast operations on a BFS tree. The most involved part is multiplying the (implicitly constructed) congestion approximator R with an arbitrary demand vector \mathbf{b} , and multiplying the transposed of the approximator matrix, R^T , with a given vector that specifies a *cost* for each edge of the trees.

Multiplying by R is done by exploiting that routing on trees is trivial and using standard techniques: during the construction, we already decomposed each tree into $O(\sqrt{n})$ components of strong diameter $O(\sqrt{n})$, which can be used to solve partially by contracting components, making the resulting tree of $O(\sqrt{n})$ nodes globally known, then determine modified demand vectors for the components out of the now locally computable partial solution, and finally resolve these remaining demands within each component. Multiplication with R^\top is implemented using similar ideas. We refer to [Section 9](#) for a detailed discussion of these procedures. Plugging the building blocks outlined in this section into this machinery, we obtain our main result [Theorem 1.1](#).

4 Outline of Distributed Congestion Approximator Construction

In this section, we outline how to adapt Madry’s construction to its recursive application in the distributed setting. In [Section 8](#), we formally prove that we achieve the same guarantees as Madry’s distribution [19] in each recursive step and that our distributed implementation is fast. Here, we focus on presenting the main ideas of the required modifications to Madry’s scheme and its distributed implementation; to this end, it suffices to consider the construction of a single step of the recursion.

Centralized Algorithm. As a starting point, let us summarize the main steps of one iteration of the centralized construction. We state a slightly simplified variant of Madry’s construction, which offers the same worst-case performance and is a better starting point for what follows. From the previous step of constructing the distribution, an edge length function ℓ_e is known (in the distributed setting, this knowledge will be local). Given $j \leq n - 1$, the following construction yields a $\Theta(j)$ -tree.

1. Compute a spanning tree \mathcal{T} of G of stretch α .
2. For each edge $e = \{v, w\} \in E$ of the graph G , route $\text{cap}(e)$ units of a commodity com_e from v to w on (the unique path from v to w in) \mathcal{T} .⁴ Denote by \mathbf{f} the vector of the sum of absolute flows passing through the edges of \mathcal{T} . Recall that $\max_{e \in E} \{\text{cap}(e)\} \in \text{poly } n$ and thus $\|\mathbf{f}\|_\infty \in \text{poly}(n)$.
3. For $e \in \mathcal{T}$, define the *relative load* of e as $\text{rload}(e) := |f_e|/\text{cap}(e) \in \text{poly } n$. We decompose the edge set of \mathcal{T} into $O(\log n)$ subsets $\mathcal{F}_i, i \in \{1, \dots, \lceil \log(\|\mathbf{f}\|_\infty + 1) \rceil\}$, where $e \in \mathcal{T}$ is in \mathcal{F}_i if $\text{rload}(e) \in (R/2^i, R/2^{i-1}]$ for $R := \max_{e \in \mathcal{T}} \{\text{rload}(e)\}$. As \mathcal{T} has $n - 1 \geq j$ edges, there must be some \mathcal{F}_i with $\Omega(j/\log n)$ edges; let i_0 be minimal with this property. Define $\mathcal{F} := \{e \in \mathcal{T} \mid \text{rload}(e) > 2^{i_0-1}\}$. Note that $|\mathcal{F}| \leq j$.
4. $\mathcal{T} \setminus \mathcal{F}$ is a spanning forest of at most $j + 1$ components. Define H as the graph on node set V whose edge set is the union of $\mathcal{T} \setminus \mathcal{F}$ and all edges of G between different components of $(V, \mathcal{T} \setminus \mathcal{F})$.
5. For components C and C' of $(V, \mathcal{T} \setminus \mathcal{F})$, pick arbitrary $v \in C$ and $w \in C'$ and denote by $p(C, C') \in C$ the last node from C on the v - w path in \mathcal{T} ; note that $p(C, C')$ does not depend on the choice of v and w . Denote by P the set of such *portals*. Replace all edges between different components C, C' of $(V, \mathcal{T} \setminus \mathcal{F})$ by parallel edges $\{p(C, C'), p(C', C)\}$ (of the same weight).
6. In the resulting multigraph, iteratively delete nodes from $V \setminus P$ of degree 1 until no such node remains. Note that the leaves of the induced subtree of \mathcal{T} must be in P , showing that the number of remaining nodes in $V \setminus P$ of degree larger than 2 is bounded by $|P| - 1 < 2j$. Add all such nodes to P .
7. For each path with endpoints in P and no inner nodes in P , delete an edge of minimum capacity and replace it by an edge of the same capacity between its endpoints.
8. Re-add the nodes and edges of $\mathcal{T} \setminus \mathcal{F}$ that have been deleted in Step 6.
9. For any $p, q \in P$, merge all parallel edges $\{p, q\}$ into a single one whose capacity is the sum of the individual capacities. The result is a j' -tree for $j' = |P| < 4j$.

⁴The difference to a single commodity is simply that flows in opposing directions do not cancel out. This means that *any* given feasible (i.e., congestion-1) flow in G can be routed on T with at most the congestion of this multi-commodity flow.

In his paper, Madry provides a scheme for updating the edge lengths between iterations so that this construction results in a distribution on $\tilde{O}(m/j)$ $\Theta(j)$ -trees that approximate cuts up to an expected $O(\alpha)$ -factor, where α is the stretch of the spanning tree construction. Updating the edge length function poses no challenges, so we will focus on the distributed implementation of the above steps in this section.

Differences to the Centralized Algorithm. Before we come to the distributed algorithm, let us first discuss a few changes we make to the algorithm in centralized terms. These do not affect the reasoning underlying the scheme, but greatly simplify its distributed implementation.

- We will omit the last step of the algorithm and instead operate on cores that are multigraphs. This changes the computed distribution, as we formally use a different graph as input to the recursion. However, Räcke’s arguments (and Madry’s generalization) work equally well on multigraphs, as one can see by replacing each edge of the multigraph by a path of length 2, where both edges have the same capacity as the original edge. This recovers a graph of $2m$ edges from a multigraph of m edges without affecting the cut structure, and the resulting trees can be interpreted as trees on the multigraph by contraction of the previously expanded edges. Similarly, both the low average stretch spanning tree construction and the cut sparsifier work on multigraphs without modification.
- After computing the spanning tree, we will immediately delete a subset of $\tilde{O}(\sqrt{n})$ edges to ensure that the new clusters will have low-depth spanning trees. The deleted edges are replaced by all edges of G crossing the corresponding cuts and will end up in the core. The same procedure is, in fact, applied to all edges selected into \mathcal{F} in Step 3 of the centralized routine; Madry’s arguments show that removing *any* subset of edges of \mathcal{T} and replacing it this way can only improve the quality of cut approximation. The main point of his analysis is that choosing \mathcal{F} in the way he does guarantees that, in terms of constructing the final distribution of j -trees, progress proportional to the number of edges in \mathcal{R}_{i_0} is made. We will apply the construction to cores of size $n' \gg \tilde{O}(\sqrt{n})$, which implies that removing the additional edges has asymptotically no effect on the progress guarantee.
- In the counterpart to Step 6 in Madry’s routine, also nodes from P may be removed if their degree becomes 1. Also here, there is no asymptotic difference in the worst-case performance of our routine from Madry’s.

To simplify the presentation, in this section we will assume that all trees involved in the construction have depth $\tilde{O}(\sqrt{n})$. This means that we can omit the deletion of $\tilde{O}(\sqrt{n})$ additional edges and further related technicalities. The general case is handled by standard techniques for decomposing trees into $O(\sqrt{n})$ components of depth $\tilde{O}(\sqrt{n})$ and relying on a BFS tree to communicate “summaries” of the components to all nodes in the graph within $\tilde{O}(\sqrt{n} + D)$ rounds (full details are given in [Section 8](#)). This approach was first used for MST construction [[18](#)]; we use a simpler randomized variant (cf. [Lemma 8.2](#)).

Cluster Graphs. Recall that we will recursively call (a variant of) the above centralized procedure on the core. We need to simulate the algorithm on the core by communicating on G . To this end, we will use *cluster graphs* (see [Section 5](#)), in which G is decomposed into components that play the role of core nodes. We will maintain the following invariants during the recursive construction:

1. There is a one-to-one correspondence between core nodes and *clusters*.
2. Each cluster c has a rooted spanning tree of depth $\tilde{O}(\sqrt{n})$.
3. No other edges exist inside clusters. Contracting clusters yields the multigraph resulting from the above construction without Step 9. From now on, we will refer to this multigraph as the core.
4. All edges in the (non-contracted) graph are also edges of G , and their endpoints know their lengths from the previous.

Overview of the Distributed Routine. We follow the same strategy as the centralized algorithm, with the modifications discussed above. This implies that the core edges for the next recursive call will simply be the graph edges between the newly constructed clusters. The following sketches the main steps of the

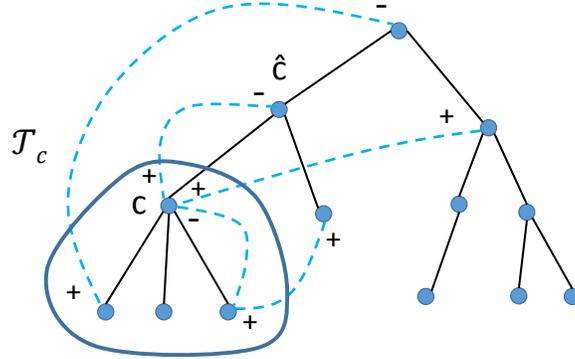


Figure 2: Illustration of the underlying idea of the aggregation scheme for the cut capacities. The cut corresponding to edge (c, \hat{c}) of the tree has a total capacity given by all graph edges leaving the subtree \mathcal{T}_c . By labeling the endpoint of graph edge by “+” if it leaves the subtree and by “-” if it connects to a descendant, the cut capacity is thus the sum of all capacities of edges labeled “+” minus all those of edges labeled “-” within \mathcal{T}_c .

distributed implementation of our overall approach.

1. Compute a spanning tree \mathcal{T} of stretch α of the core. This is done by the spanning tree algorithm of [Theorem 3.1](#), which can operate on the cluster graph.
2. For each edge $e \in \mathcal{T}$, determine its absolute flow $|f_e|$ (and thus $\text{rload}(e) = |f_e|/\text{cap}(e)$) as follows (cf. [Figure 2](#)).
 - (*) For each cluster c , consider the cut induced by the edge to its parent. For each “side” of the cut, we want to determine the total capacity of all edges incident to nodes of c that connect to the respective side of the cut. Denote by c_+ the total “outgoing” capacity of cluster c towards the root’s side and by c_- the “incoming” capacity.
 - (a) Each cluster c learns its ancestor clusters in the spanning tree of C .
 - (b) Observe that for a cluster c , an edge does contribute to c_- if and only if it connects to a node within its subtree \mathcal{T}_c . From the previous step, this information is known to one of the endpoints of the edge. We communicate this and determine in each cluster c the values c_+ and c_- by aggregation on its spanning tree.
 - (c) Suppose $e \in \mathcal{T}$ is the edge from cluster c to its parent. Using aggregation on the spanning tree of C , we compute

$$|f_e| = \sum_{c' \in \mathcal{T}_c} c'_+ - c'_-.$$

3. Determine the index i_0 (as in Step 3 of the centralized routine). Given that $\text{rload}(e)$ for each $e \in \mathcal{T}$ is locally known, this is performed in $\tilde{O}(D)$ rounds using binary search in combination with converge-and-broadcasts on a BFS tree. We set $\mathcal{F} := \{e \in \mathcal{T} \mid \text{rload}(e) > 2^{i_0-1}\}$.
4. Define P as the set of clusters incident to edges in \mathcal{F} . A simple broadcast on the cluster spanning trees makes membership known to all nodes of each cluster $c \in P$.
5. Iteratively mark clusters $c \notin P$ with at most one unmarked neighboring cluster, until this process stops. Add all unmarked clusters that retain more than 2 unmarked neighboring clusters to P .
6. For each path with endpoints in P whose inner nodes are unmarked clusters not in P , find the edge $e \in \mathcal{T} \setminus \mathcal{F}$ of minimal capacity and add it to \mathcal{F} . This disconnects any two clusters $c, c' \in P$, $c \neq c'$, in $\mathcal{T} \setminus \mathcal{F}$.
7. Each component of $\mathcal{T} \setminus \mathcal{F}$ and the spanning trees of clusters induce a spanning tree of the corresponding component of G . Each such component is a new cluster. Make the identifier of the unique $c \in P$ of each cluster known to its nodes and delete all edges between nodes in the cluster that are not part of its spanning tree.

If all trees have depth $\tilde{O}(\sqrt{n})$, all the above steps can be completed in $\tilde{O}(\sqrt{n} + D)$ rounds. Clearly, the first 3 stated invariants are satisfied by the given construction. As mentioned earlier, it is also straightforward to update the edge lengths, i.e., establish the fourth invariant. Once the distribution on the current level of recursion is computed, one can hence sample and then move on to the next level.

For the detailed description of the algorithm, the recursion, a formal statement of [Theorem 3.2](#), and the respective proofs, we refer to [Section 8](#).

5 Cluster Graphs

On several levels, our distributed congestion approximator construction is done in a hierarchical way. As a consequence many of the distributed computations used by our algorithm have to be run on a graph induced by clusters of the network graph. In order to be able to deal with such cluster graphs in a systematic way, we formally define cluster graphs and we describe how to simulate distributed computations on a cluster graph by running a distributed algorithm on the underlying network graph.

Definition 5.1 (Distributed Cluster Graph). *Given a n -network graph $G = (V, E)$, a distributed N -node cluster graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{L}, \mathcal{T}, \psi)$ of size n is defined by a set of N clusters $\mathcal{V} = \{S_1, \dots, S_N\}$ partitioning the vertex set V , a set (or multi-set) of edges $\mathcal{E} \subseteq \binom{\mathcal{V}}{2}$, a set of cluster leaders \mathcal{L} , a set of cluster trees \mathcal{T} , as well as a function ψ that maps the edges \mathcal{E} of the cluster graph to edges in E . Formally, the tuple $(\mathcal{V}, \mathcal{E}, \mathcal{L}, \mathcal{T}, \psi)$ has to satisfy the following conditions.*

- (I) *The clusters $\mathcal{V} = (S_1, \dots, S_N)$ form a partition of the set of vertices V of the network graph, i.e., $\forall i \in [N] : S_i \subseteq V, \forall 1 \leq i < j \leq N : S_i \cap S_j = \emptyset$, and $\bigcup_{i=1}^N S_i = V$.*
- (II) *For each cluster S_i , $|S_i \cap \mathcal{L}| = 1$. Hence, each cluster has exactly one cluster leader $\ell_i \in \mathcal{L} \cap S_i$. The ID of the node ℓ_i also serves as the ID of the cluster S_i and for the purpose of distributed computations, we assume that all nodes $v \in S_i$ know the cluster ID and the size $n_i := |S_i|$ of their cluster S_i .*
- (III) *Each cluster tree $T_i = (S_i, E_i)$ is a rooted spanning tree of the subgraph $G[S_i]$ of G induced by S_i . The root of T_i is the cluster leader $\ell_i \in S_i \cap \mathcal{L}$. We assume that each node of $u \in S_i \setminus \{\ell_i\}$ knows its parent node $v \in S_i$ in the tree T_i .*
- (IV) *The function $\psi : \mathcal{E} \rightarrow E$ maps each edge $\{S_i, S_j\} \in \mathcal{E}$ to an (actual) edge $\{v_i, v_j\} \in E$ connecting the clusters S_i and S_j , i.e., it holds that $v_i \in S_i$ and $v_j \in S_j$. The two nodes v_i and v_j know that the edge $\{v_i, v_j\}$ is used to connect clusters S_i and S_j . If the cluster graph is weighted, the two nodes v_i and v_j also know the weight of the edge $\{S_i, S_j\}$.*

Note that (III) in particular implies that the subgraph of G induced by each cluster S_i is connected. When dealing with a concrete distributed cluster graph \mathcal{G} , we use $\mathcal{G}_{\mathcal{V}}, \mathcal{G}_{\mathcal{E}}, \mathcal{G}_{\mathcal{L}}, \mathcal{G}_{\mathcal{T}}$, and \mathcal{G}_{ψ} to denote the corresponding sets of clusters, edges, etc. Further, when only arguing about the cluster graph and not its mapping to G , we only use the pair $(\mathcal{V}, \mathcal{E})$ to refer to it. In the following, we say that a cluster $S \in \mathcal{V}$ knows something if all nodes $v \in S$ know it. That is, e.g., the last part of condition (II) says that every cluster knows its ID and its size.

We next define a weak version of the (synchronous) CONGEST model and we show that algorithms in this model can be efficiently simulated in distributed cluster graphs.

Definition 5.2 (B -Bounded Space CONGEST Model). *Let $B = \Omega(\log n)$ be a given parameter. The B -Bounded Space CONGEST model is a computation model which restricts the CONGEST(B) model by requiring, for any $d \geq 0$, that each step of a node v in the B -Bounded Space can be emulated in $O(d)$ rounds by any tree $T(v)$ of depth d , where the edges incident on v are incident on nodes of $T(v)$ in the emulation.*

The definition of the B -Bounded Space model is directed toward emulation. The definition immediately implies that if each node is emulated by a tree, then emulating a global step in time proportional to the maximal tree depth (which could be $\Omega(n)$) is trivial. However, the following lemma shows that this can actually be done in time $O(D + \sqrt{n})$.

Lemma 5.1. *Given an underlying n -node graph $G = (V, E)$ and a cluster graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{L}, \mathfrak{T}, \text{psi})$, a t -round distributed algorithm \mathcal{A} in \mathcal{G} in the B -bounded space CONGEST model can be simulated in the (ordinary) CONGEST model in G with messages of size at most B in $O((D + \sqrt{n}) \cdot t)$ rounds, where D is the diameter of G .*

Proof Sketch. We assume that we are given a global BFS tree of G . If such a BFS tree is not available, it can be computed in $O(D)$ rounds in the CONGEST model. We simulate the algorithm \mathcal{A} in a round-by-round manner. Consider the end of the simulation of round $r - 1$ and assume that in each cluster $S_i \in \mathcal{V}$, the leader node ℓ_i knows the message M_i to be sent in round r . (For $r = 1$, we assume that this is true at the beginning of the simulation.)

To start the simulation of round r , we first make sure that for every $S_i \in \mathcal{V}$, every node $v \in S_i$ knows the message M_i to be sent to the neighbors in round r . In clusters S_i of size at most \sqrt{n} , this can be done in at most \sqrt{n} rounds by broadcasting M_i on the spanning tree T_i of $G[S_i]$. For larger clusters, we use the global BFS tree to disseminate the information. We first send all the messages M_i of clusters S_i of size larger than \sqrt{n} to the root of the global BFS tree. Because the BFS tree has radius at most D and because there are at most \sqrt{n} clusters of size larger than \sqrt{n} , this can be done in $D + \sqrt{n}$ rounds (using pipelining). Now, in another $D + \sqrt{n}$ rounds, all these messages can be broadcast to all nodes of G (and thus also to the nodes of the clusters that need to know them).

Now, for every two clusters S_i and S_j such that $\{S_i, S_j\} \in \mathcal{E}$, let $\{u_i, u_j\} = \text{psi}(\{S_i, S_j\})$ be the physical edge connecting S_i and S_j . The node $u_i \in S_i$ sends the message M_i to u_j and the node $u_j \in S_j$ sends M_j to u_i . This step can be done in a single round. Now, in each cluster S_i , each incoming message of round r is known by one node in S_i and we need to aggregate these messages in order to compute the outgoing message of each cluster. In clusters of size at most \sqrt{n} , this can again be done locally inside the cluster (by Definition 5.2). Also, for the at most \sqrt{n} clusters of size larger than \sqrt{n} , we again use the global BFS tree. Since in a tree of depth D , k independent convergcasts of broadcasts can be done in time $D + k$, the messages of the large clusters can be computed in disseminated to the cluster leaders in time $O(D + \sqrt{n})$. \square

6 Distributed Construction of Cut Sparsifiers

Lemma 6.1. *In a weighted N -node distributed cluster graph of size n , for any $\varepsilon > 0$, it is possible to compute a spectral $(1 + \varepsilon)$ -sparsifier with $O(N \cdot (\varepsilon^{-1} \cdot \log N)^{O(1)})$ edges (w.h.p.) in the CONGEST model in time $O((D + \sqrt{n}) \cdot (\varepsilon^{-1} \cdot \log N)^{O(1)})$. When the algorithm terminates, each of the edges of the sparsifier is directed such that the out-degree of each cluster is upper bounded by $O((\varepsilon^{-1} \cdot \log N)^{O(1)})$ and such that each cluster knows all its outgoing edges.*

Proof. We prove the lemma using the algorithm PARALLELSPARSIFY of Koutis [17], and then orient the edges. Koutis's algorithm relies on the $O(\log n)$ -stretch spanner construction algorithm of Baswana and Sen [10], which we henceforth refer to as BS. See Figure 3 for a description of the BS algorithm.

We start by showing how to emulate a step of node v in BS by a depth- d tree T_v in $O(d + \log N)$ time (w.h.p.). We may assume w.l.o.g. the existence of a root in each tree (because we can select one in $O(d)$ time); Step 2a is carried out by the root and the result is broadcast over the tree. For Step 2b, we note that w.h.p., $|Q_v| = O(\log N)$ and hence making it known to all node T_v takes $O(d + \log N)$ time using standard convergcast-broadcast. Step 3 is straightforward given that each node knows its cluster.

1. $R_0 := \{\{v\} \mid v \in V\}$.
2. For $i := 1$ to $\log N$ do:
 - (a) Mark each cluster of R_{i-1} independently with probability $\frac{1}{2}$; let $R_i := \{S \text{ is marked} \mid S \in R_{i-1}\}$.
 - (b) If $v \in S$ for some $S \in R_{i-1} \setminus R_i$:
 - i. Define Q_v to be the set of edges that consists of the lightest edge from v to each of the clusters in R_i v is adjacent to.
 - ii. If v has no neighbor in a cluster in R_i , then v adds to the spanner all edges in Q_v .
 - iii. Otherwise, let u be the closest neighbor of v in a marked cluster. Then
 - v joins the cluster of u (i.e., if u is in cluster $S' \in R_i$, then $S' := S' \cup \{v\}$).
 - v adds to the spanner the edge $\{v, u\}$, and also all edges $\{v, w\} \in Q_v$ with $W(v, w) < W(v, u)$ (breaking ties by ID).
3. Each node v adds, for each cluster $S \in R_{\log N}$ it is adjacent to, the lightest edge connecting it to S .

Figure 3: The BS algorithm for $O(\log N)$ spanner construction given an N -node weighted graph $G = (V, E, W)$. The output is a subset of E .

Next, given a tree $T(v)$ for each node, we assume that the depth of all trees is at least $c \log N$ for some appropriate constant $c > 0$. If this assumption does not hold we extend $T(v)$ with a dummy path: clearly $T(v)$ can emulate the extended tree without any slowdown. However this extension may increase the number of nodes by an $O(\log N)$ factor. Now, under this assumption and the emulation above, we may apply Lemma 5.1 to conclude that BS can be executed in time $O((D + \sqrt{N \log N}) \log N)$.

Going back to the algorithm of Koutis [17], we note that it consists of $(\log n/\varepsilon)^{O(1)}$ invocations of BS, and some independent random selection and reweighting of edges. The former is discussed above, and the latter is trivial to emulate locally.

Finally, for edge orientation, we give a little algorithm that, given an N -node, D -diameter graph with average degree d_{av} , orients all edges such that the out degree of all nodes is $O(d_{av})$. The algorithm runs in $O(D + \log n)$ steps in the space-bounded CONGEST model. The algorithm is as follows. First compute the average degree in $O(D)$ time, and then repeat the following procedure $\log n$ times at each node v :

- If the number of unoriented edges incident on v is less than $2d_{av}$, then v orients all unoriented edges outward, informs its neighbors, and halts.

The correctness of the procedure follows from the fact that throughout the execution, at most half the non-halted nodes have degree larger than $2d_{av}$. The lemma now follows from the fact that the graph generated by Koutis' algorithm has average degree $(\frac{\log N}{\varepsilon})^{O(1)}$. \square

7 Distributed Construction of Low Average-Stretch Spanning Trees

Theorem 3.1. (restated and rephrased) Suppose H is a multigraph obtained from G by assigning arbitrary edge lengths in $2^{n^{o(1)}}$ to the edges of G (known to incident nodes) and performing an arbitrary sequence of contractions. Then we can compute a rooted spanning tree of H of expected stretch $2^{O(\sqrt{\log n \log \log n})}$ within $(\sqrt{n} + D)n^{o(1)}$ rounds, where the edges of the tree in H and their orientation is locally known to the endpoints of the corresponding edges in G .

Proof. We follow [12]: the high-level algorithm is by Alon et al. [3], which uses Algorithm Partition (of [12]) for unweighted graphs. We describe the algorithm bottom-up. The main component in Algorithm Partition is Algorithm SplitGraph, reproduced in Figure 4. The basic action of Algorithm SplitGraph is growing BFS trees, an action in which emulating a single node by a tree is trivial. In SplitGraph we may have contending BFS growths, but note that if two or more BFS traversals collide, only the winning ID

1. $G^1 = (V^1, E^1) := G; \mathcal{C} := \emptyset$.
2. For $t = 1$ to $2 \log N$ do:
 - (a) Let S^t random subset of V^t of $12 \frac{2^{t/2}}{n} |V^t|$ nodes; if V^t is smaller than $\frac{n}{12 \cdot 2^{t/2}}$, $S^t := V^t$.
 - (b) $\mathcal{C} := \mathcal{C} \cup \{\{s\} \mid s \in S^t\}$.
 - (c) Each $s \in S^t$ draws a random delay δ_s^t uniformly from $[0, \lfloor \rho / (2 \log N) \rfloor]$.
 - (d) Each $s \in S^t$ waits δ_s^t rounds and then initiates a BFS for $\rho(1 - \frac{t-1}{2 \log N}) - \delta_s^t$ rounds in G^t .
 - (e) A node covered by a BFS is added to cluster C_s , where s is the source of the first BFS to visit it, breaking ties by ID.
 - (f) $V^{t+1} := V \setminus \{v \mid v \in C \text{ for some } C \in \mathcal{C}\}; G^{t+1} := G^t[V^{t+1}]$.

Figure 4: Algorithm SplitGraph. The input is an unweighted graph $G = (V, E)$ and a target radius ρ .

needs to proceed, and hence there are no collisions because no edge needs to carry more than a single BFS traversal in each direction. Regarding the tree construction, we first note that the BFS growth naturally creates a spanning tree for each cluster. Moreover, we can make all nodes know the complete path to the root of their respective cluster in additional $O(\rho)$ steps, by letting each node send its i^{th} ancestor to all its children in round i . The running time of Algorithm SplitGraph is clearly $O(\rho \log N)$ in the bounded-space CONGEST model, and therefore, using Lemma 5.1, we conclude that we can run SplitGraph in time $O(\rho \log N(D + \sqrt{N}))$ in the CONGEST model.

Algorithm SplitGraph is called by Algorithm Partition, whose input is an unweighted graph with an arbitrary partition of the edges into K classes. Algorithm Partition applies Algorithm SplitGraph disregarding classes, and then checks whether there exists a class where too many edges were split in different clusters. If there is such an over-split class, the algorithm is restarted. We can implement each checking and restart in the CONGEST model in $O(D + K)$ time using a global BFS tree. Since the number of restarts is bounded by $O(\log N)$ w.h.p. [12], and in our implementation we shall have $K = O(\sqrt{N})$, the overall time for running Algorithm Partition is $O(\rho \log^2 N(D + \sqrt{N}))$ in the CONGEST model.

The outermost algorithm is the one by Alon et al. [3], whose input is a *weighted* graph. The algorithm first partitions the edges into $O(\sqrt{\log N})$ classes by weight, where class E_i contains all edges whose weight is in $[z^{i-1}, z^i)$ for a certain value $z = \tilde{\Theta}(2^{\sqrt{6 \log N \cdot \log \log N}})$. Then the algorithm proceeds in iterations until the graph is a single node, where iteration j is as follows.

1. Call Algorithm Partition with edges E_1, \dots, E_j and target radius $\rho = z/4$. Obtain clusters $\{C_i\}$.
2. Output a BFS tree for each cluster C_i .
3. Contract each resulting cluster C_i to a single node. Remove all self loops, but leave parallel edges in place. The resulting multigraph, augmented with edge class E_{j+1} , is the input to iteration $j + 1$.

For the distributed implementation, note that edge contraction is trivial given that the endpoints know the identity of the cluster they belong to, and that edge classification is purely local given z (which can be communicated to all in $O(D)$ time units). It can be shown [12] that w.h.p., the number of iterations is $O(\log \Delta / \sqrt{\log N \log \log N})$, and hence the running time of the algorithm is $O(\rho \log \Delta \log^{O(1)} N(D + \sqrt{N})) = \log \Delta \cdot 2^{O(\sqrt{\log N \log \log N})}$ because $\rho = \tilde{\Theta}(2^{\sqrt{6 \log N \cdot \log \log N}})$.

The claimed stretch follows from [12]. □

8 Distributed Construction of j -Trees

From the distributed implementation point of view, the core technical challenge is to efficiently compute a congestion approximator in a distributed way. As already pointed out, the congestion approximator is constructed based on applying the j -tree construction of Madry [19] recursively. In the following, we review Madry's construction and we show how to implement an adapted version of it in a distributed network. The

main objective of the construction is to approximate the flow structure of a given graph by a distribution of graphs from a simpler class of graphs (i.e., j -trees). Formally, the similarity of the flow structure of two graphs is captured by the following definition from [19].

Definition 8.1 (Graph Embeddability). [19] We are given $\beta \geq 1$ and two (multi-)graphs $G = (V, E, \text{cap})$ and $G' = (V, E', \text{cap}')$ on the same set of nodes and with edge capacities $\text{cap}(e)$ and $\text{cap}'(e')$ for edges $e \in E$ and $e' \in E'$. We say that graph G is β -embeddable into G' if there exists a multicommodity flow $\mathbf{f}' = (\mathbf{f}'_e)_{e \in E}$ such that for every edge $e \in E$ of G connecting nodes u and v , \mathbf{f}'_e is a flow on $(V, E', \beta \text{cap}')$ that routes $\text{cap}(e)$ units of flow between u and v , and for every edge $e' \in E'$ of G' , it holds that $|\mathbf{f}'(e')| := \sum_{e \in E} |(\mathbf{f}'_e)(e')| \leq \beta \text{cap}'(e')$.

Intuitively, a graph G is β -embeddable into a graph G' , if for every (multicommodity) flow problem, there is a solution in G' such that the maximum relative congestion of all edges is by at most a factor β larger than for the optimal solution in G . As a generalization of the cut-based graph decompositions of Räcke [26], Madry defines the notion of an (α, \mathbb{G}) -decomposition.

Definition 8.2 (α, \mathbb{G}) -Decomposition [19]. Given a (multi-)graph $G = (V, E, \text{cap})$ and a family \mathbb{G} of graphs on the nodes V , an (α, \mathbb{G}) of G is a set of pairs $\{(\lambda_i, G_i)\}_{i \in I}$ satisfying that:

- $\forall i \in I : \lambda_i > 0$;
- $\sum_{i \in I} \lambda_i = 1$;
- $\forall i \in I : G_i = (V, E_i, \text{cap}_i)$ is a graph in \mathbb{G} ;
- $\forall i \in I : G$ is 1-embeddable into G_i ; and
- the graph defined by the convex combination⁵ $\sum_{i \in I} \lambda_i \cdot G_i$ is α -embeddable into G .

In words, $\{(\lambda_i, G_i)\}_{i \in I}$ is a distribution on I graphs from \mathbb{G} , each of which can be 1-embedded into G , such that the distribution α -embeds into G .

Observe that such a decomposition can form the basis for a good congestion approximator: 1-embeddability of each G_i into G guarantees that congestion is never overestimated, and the embeddability of the convex combination ensures when sampling from the distribution, the expected factor by which we underestimate congestion on a cut is at most α . Our goals are now to choose \mathbb{G} and the distribution such that

- α is small,
- we can construct the distribution efficiently, and
- we can evaluate the induced congestion when routing demand optimally on a graph from the distribution efficiently.

The Plan. Let $G = (V, E, \text{cap})$ be a weighted (multi-)graph, $0 \leq j \leq |V|$ be an integer and let \mathbb{J} be the family of j -trees over the node set V . In [19], it is shown that based on a protocol for computing spanning trees with average stretch α , there exists an (α, \mathbb{J}) -decomposition of G . This is shown in several steps. It is first shown that a sparse (α, \mathbb{H}) -decomposition exists for a graph family \mathbb{H} which contains graphs that are closer to the original graph G and it is then shown that every graph $\mathcal{H} \in \mathbb{H}$ can be $O(1)$ -embedded into a j -tree and vice versa.

As described, we have to apply the j -tree construction recursively to the core graph. Each node in the core graph is represented by a set of nodes (a cluster) in the network graph. On the network graph, the core graph therefore corresponds to a graph between clusters of nodes. We therefore have to be able to apply the j -tree construction on a cluster graph. As we will see, we can construct j -trees such that whenever two nodes u and v of the core are connected by a (virtual) edge, there also is a physical edge between the two trees (i.e., clusters of nodes) corresponding to u and v . Throughout our algorithm, we can therefore work

⁵The sum of two weighted graphs $G_1 = (V, E_1, \text{cap}_1)$ and $G_2 = (V, E_2, \text{cap}_2)$ is defined as $G_1 + G_2 = (V, E_1 \cup E_2, \text{cap}_{12})$, where for each edge $e \in E_1 \cup E_2$, $\text{cap}_{12}(e)$ is defined as $\text{cap}_{12}(e) = \text{cap}_1(e) + \text{cap}_2(e)$ if $e \in E_1 \cap E_2$ and $\text{cap}_{12}(e) = \text{cap}_i(e)$ if $e \in E_i \setminus E_{3-i}$ for $i \in \{1, 2\}$.

with a cluster multigraph such that a) the induced graph of each cluster is connected and b) for every edge between two clusters c and c' , there are nodes $u \in c$ and $v \in c'$ such that u and v are connected by an edge in the underlying network graph. For doing distributed computations, we assume that each cluster has a leader and that every node knows the ID of the leader and also its parent in a rooted spanning tree which is rooted at the leader. In [Section 5](#), we give a precise definition of a *distributed cluster graph* and we show that several basic algorithms that we use as building blocks can be run efficiently in distributed cluster graphs.

In the following, we go through Madry's j -tree construction step-by-step and describe how to adapt it so that we can implement it efficiently on a distributed cluster graph (i.e., in the CONGEST model in the underlying network graph).

8.1 Low-Stretch Spanning Trees

In the following, we consider the computation of the (α, \mathbb{J}) -decomposition of some core graph. Assume that the core graph is given as a distributed cluster graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \text{cap})$, where for each edge $e \in \mathcal{E}$, $\text{cap}(e)$ is the capacity of e . As the time complexity of some of the steps for computing an (α, \mathbb{J}) -decomposition of \mathcal{G} depend on the number of edges of \mathcal{G} , as a first step, we sparsify \mathcal{G} . In [Lemma 6.1](#), it is shown that in $O((D + \sqrt{n}) \cdot \text{polylog } n)$ rounds, it is possible to compute an $(1 + 1/\text{polylog } n)$ -spectral sparsifier of \mathcal{G} with at most $O(|\mathcal{V}| \cdot \text{polylog } n)$ edges. Further, for each edge $\{c, c'\} \in \mathcal{E}$ of the sparsifier one of the nodes of the edge manages the edge. As in general \mathcal{G} is a cluster graph, c and c' are clusters of physical nodes and an edge connecting clusters c and c' is represented by a physical edge $\{u, v\}$ for two (network) nodes $u \in c$ and $v \in c'$. We will maintain that every pair of nodes $u \in c$ and $v \in c'$ needs to represent at most one edge between c and c' in \mathcal{G} . The two nodes u and v know about the edge between c and c' and its capacity.

In the following, we assume that \mathcal{G} is the graph after sparsification. If the number of nodes $|\mathcal{V}|$ of \mathcal{G} is less than $n^{1/2+o(1)}$, using a global BFS tree of the network graph, the whole structure of \mathcal{G} can be collected in $O(D + |\mathcal{V}| \text{polylog } n) = O(D + n^{1/2+o(1)})$ rounds. In that case, we can therefore perform all remaining operations locally at the nodes. Consequently, we will henceforth assume that $|\mathcal{V}| \geq n^{1/2+o(1)}$.

During the construction of the (α, \mathbb{J}) -decomposition of \mathcal{G} , each edge $e \in \mathcal{E}$ is assigned a length $\ell(e)$. At the beginning $\ell(e)$ is proportional to $1/\text{cap}(e)$ and before adding each j -tree, $\ell(e)$ is adapted for each edge. As the first step of constructing each j -tree in the decomposition, Madry computes a spanning tree \mathcal{T} of \mathcal{G} for which it holds that

$$\sum_{e=\{c,c'\} \in \mathcal{E}} d_{\mathcal{T}}(c, c') \cdot \text{cap}(e) = \sum_{e=\{c,c'\} \in \mathcal{E}} \text{stretch}_{\mathcal{T}}(e) \cdot \ell(e) \cdot \text{cap}(e) \leq \delta \alpha \cdot \sum_{e=\{c,c'\} \in \mathcal{E}} \ell(e) \cdot \text{cap}(e) \quad (2)$$

for a sufficiently small positive constant δ . In the above expression, $d_{\mathcal{T}}(u, v)$ denotes the sum of edge lengths on the path between u and v on \mathcal{T} . Hence, \mathcal{T} is a spanning tree with a bounded weighted average stretch. Such a spanning tree can be computed by computing an (unweighted) low average stretch spanning tree for a multigraph $\tilde{\mathcal{G}}$ which is obtained from \mathcal{G} by (logically) replacing some of the edges of \mathcal{G} with multiple copies of the same edge (overall, the number of edges is as most doubled) [[4](#), [19](#)].

In our distributed implementation of Madry's j -tree construction, we adapt the parallel low average stretch spanning tree algorithm from [[12](#)] to our setting. The algorithm of [[12](#)] already works in a mostly decentralized fashion and we can therefore also apply it in a distributed setting. In [Section 7](#), we show how to run the algorithm of [[12](#)] on a distributed cluster graph. We note that the low average stretch spanning tree algorithm of [[12](#)] directly tolerates multi-edges as described above, even if the same physical edge has to be used to represent multiple edges between the same clusters.

Theorem 3.1. *(restated and rephrased) Suppose \mathcal{G} is a multigraph obtained from G by assigning arbitrary edge lengths in $2^{n^{o(1)}}$ to the edges of G (known to incident nodes) and performing an arbitrary sequence of contractions. Then we can compute a rooted spanning tree of \mathcal{G} of expected stretch $2^{O(\sqrt{\log n \log \log n})}$*

within $(\sqrt{n} + D)n^{o(1)}$ rounds, where the edges of the tree in \mathcal{G} and their orientation is locally known to the endpoints of the corresponding edges in G .

Given the spanning tree $\mathcal{T} = (\mathcal{V}_{\mathcal{T}}, \mathcal{E}_{\mathcal{T}})$ of \mathcal{G} , we need to compute capacities $\text{cap}_{\mathcal{T}}(e)$ for the spanning edges such that \mathcal{G} is embeddable into \mathcal{T} , which essentially boils down to computing the absolute value of the multicommodity flow \mathbf{f}' routing $\text{cap}(e)$ units of flow on \mathcal{T} for each $e \in E$. As routing is trivial in trees, \mathbf{f}' is unique. Once $|\mathbf{f}'|$ is computed, the edge capacities of \mathcal{T} can be chosen accordingly and it is straightforward to pick a suitable λ_i and update the length function $\ell(e)$. However, computing the absolute value of the multicommodity flow *fast* in the distributed setting requires some work.

Computing the Multicommodity Flow. In the following, assume that the edges of \mathcal{T} are oriented towards the root, i.e., we will write $(c, \hat{c}) \in \mathcal{T}$ if $\hat{c} \in \mathcal{V}$ is the parent of cluster $c \in \mathcal{V}$. Denote by \mathcal{T}_c the subtree of \mathcal{T} rooted at c . When embedding \mathcal{G} into \mathcal{T} , we have to route a total of

$$|\mathbf{f}'(c, \hat{c})| = \sum_{\substack{c_1 \in \mathcal{T}_c \\ c_2 \notin \mathcal{T}_c}} \sum_{\{c_1, c_2\}_{uv} \in \mathcal{E}} \text{cap}(\{c_1, c_2\}_{uv})$$

commodity through the edge $\{c, \hat{c}\} \in \mathcal{T}$, where we indexed the different edges of the multigraph (V, \mathcal{E}) by using that for each $e \in \mathcal{E}$ between c_1 and c_2 , ψ maps e to a unique edge $\{u, v\} \in E$ with $u \in c_1$ and $w \in c_2$. Note that u and v know that $\{c_1, c_2\}_{uv} \in \mathcal{E}$, that they are in the clusters c_1 and c_2 , respectively, and what $\text{cap}(\{c_1, c_2\}_{uv})$ is. We thus have to solve the task of determining this sum for each edge $\{c, \hat{c}\} \in \mathcal{T}$ via computations on the graph G underlying $(\mathcal{V}, \mathcal{E})$.

Observe that the spanning trees of the clusters together with \mathcal{T} induce a (rooted) spanning tree T of G . Essentially, we would like to perform, for each edge $\{c, \hat{c}\} \in \mathcal{T}$, an aggregation on T and pipeline these aggregations to achieve good time complexity. However, as shown in the following lemma, the result is a running time linear in the depth of the tree, which may be $\Omega(n)$ irrespective of D .

Lemma 8.1. *If T has depth d , for each edge $e = (c, \hat{c}) \in \mathcal{T}$, c can determine $|\mathbf{f}'(e)|$ within $O(d)$ rounds.*

Proof. Consider the following algorithm.

1. For each cluster, all of its nodes learn the ancestors of the cluster in \mathcal{T} .
2. For each edge $\{c_1, c_2\}_{uv} \in \mathcal{E}$, u and v exchange the ancestor lists of c_1 and c_2 .
3. Each node $u \in c_1 \in \mathcal{T}$ locally computes for each ancestor c of c_1 the value

$$\text{cap}_c(u) := \sum_{\substack{\{c_1, c_2\}_{uv} \in \mathcal{E} \\ c \text{ is not ancestor of } c_2}} \text{cap}(\{c_1, c_2\}_{uv}).$$

4. For each edge $e = (c, \hat{c}) \in \mathcal{T}$, we aggregate $\sum_{u \in T_c} \text{cap}_c(u)$ on T_c , where T_c is the subtree of T corresponding to \mathcal{T}_c .

Observe that, by definition,

$$|\mathbf{f}'(c, \hat{c})| = \sum_{\substack{c_1 \in \mathcal{T}_c \\ c_2 \notin \mathcal{T}_c}} \sum_{\{c_1, c_2\}_{uv} \in \mathcal{E}} \text{cap}(\{c_1, c_2\}_{uv}) = \sum_{u \in T_c} \text{cap}_c(u),$$

as each edge $\{c_1, c_2\}_{uv} \in \mathcal{E}$ with $c_1 \in \mathcal{T}_c$ and $c_2 \in \mathcal{T} \setminus \mathcal{T}_c$ satisfies that either $u \in T_c$ or $v \in T_c$. Hence, it remains to show that the above routine can be implemented with a running time of $O(d)$.

Clearly, the first step takes d rounds: T has depth d , and we can perform concurrent floodings on all subtrees without causing contention. The second step requires at most $d - 1$ rounds, as no node has more than $d - 1$ ancestors. The third step requires local computations only. Finally, the fourth step can be performed in d rounds as well, since we can perform concurrent convergecasts on all subtrees without causing contention. \square

To handle the general case, i.e., $d \gg \sqrt{n}$, we first decompose \mathcal{T} into $O(\sqrt{n})$ parts of small diameter. There are different ways to achieve such a decomposition of \mathcal{T} efficiently in a distributed way (e.g., by using techniques from [18]). The easiest way is to use randomization. Suppose c' is the parent cluster of non-root cluster c . We sample edge $e = (c, \hat{c}) \in \mathcal{T}$ into the edge set \mathcal{R} with independent probability $q_e := \min\{1, |c|/\sqrt{n}\}$. Then, w.h.p. the forest $T \setminus \psi(\mathcal{R})$ consists $\tilde{O}(\sqrt{n})$ trees of depth $\tilde{O}(\sqrt{n})$.

Lemma 8.2. *Let \mathcal{T} be a rooted spanning tree of a cluster (multi-)graph \mathcal{G} and let \mathcal{R} be a subset of the edges chosen at random as described above, and assume that the spanning tree of each cluster has depth at most d . W.h.p., the forest $T \setminus \psi(\mathcal{R})$ induced by the edges $\mathcal{T} \setminus \mathcal{R}$ and the cluster spanning trees consists of $O(\sqrt{n})$ rooted trees of depth $d + O(\sqrt{n} \log n)$.*

Proof. Clearly, the number of trees in the forest induced by $\mathcal{E}_{\mathcal{T}} \setminus \mathcal{R}$ is equal to $|\mathcal{R}| + 1$. The expected value for $|\mathcal{R}|$ is given by the sum of the probabilities q_e and thus $\mathbb{E}[|\mathcal{R}|] \leq \sqrt{n}$. A standard Chernoff bound implies that $|\mathcal{R}|$ does not exceed \sqrt{n} by more than a constant factor with high probability.

To bound the depth of each (rooted) tree in $T \setminus \psi(\mathcal{R})$, consider a cluster $c \in \mathcal{T}$ and a path p from the leader $\{r\} := \mathcal{L} \cap c$ to some node in the subtree T_r of T rooted at r . The depth of $T_r \cap c$ is bounded by d . Denote by E_p the set of edges of p that correspond to edges in \mathcal{T} , i.e., each $e \in E_p$ satisfies that $e = \psi(e')$ for some $e' \in \mathcal{T}$. Denote by $c(e)$ the child cluster of e , i.e., the endpoint further away from the root of T . By construction, $c(e) \in e'$ is also the cluster further away from the root of \mathcal{T} (for the $e' \in \mathcal{T}$ with $\psi(e') = e$). Therefore, the length of p is bounded by

$$d + |E_p| + \sum_{e \in E_p} |c(e)|$$

i.e., the sum of the number of its edges in c , the number of edges between clusters $|E_p|$, and the number of edges in each traversed cluster. For $e \in E_p$, we have that $q_e = \min\{1, |c(e)|/\sqrt{n}\}$, yielding that

$$\mathbb{E}[|\mathcal{R} \cap \psi^{-1}(E_p)|] \leq \sum_{e \in E_p} \frac{|c(e)|}{\sqrt{n}}.$$

Applying Chernoff's bound shows that $\mathbb{E}[|\mathcal{R} \cap \psi^{-1}(E_p)|] \in O(\log n)$ or $|\mathcal{R} \cap \psi^{-1}(E_p)| > 0$ w.h.p. The former implies that (i) $\sum_{e \in E_p} |c(e)| \in O(\sqrt{n} \log n)$ and (ii) $|E_p| \in O(\sqrt{n} \log n)$, as always $|c(e)| \geq 1$; it follows that the length of p is bounded by $d + O(\sqrt{n} \log n)$ as claimed. The latter implies that, w.h.p., p is not contained in $T \setminus \psi(\mathcal{R})$. As the number of simple paths in a tree is bounded by $O(n^2)$, applying the union bound completes the proof. \square

For simplicity, in the following we assume that the high probability statements of the above lemma hold with certainty; the final statements then follow by applying the union bound.

Throughout the construction, we will maintain that edges of \mathcal{R} are never retained. As there will be $o(\log n)$ levels of recursion, by inductive use of the above lemma, it follows that clusters always have spanning trees of depth $\tilde{O}(\sqrt{n})$. Exploiting this property together with the small number of connected components of $\mathcal{T} \setminus \mathcal{R}$, we obtain a fast routine for the general case.

Lemma 8.3. *Within $\tilde{O}(\sqrt{n} + D)$ rounds, for each edge $e = (c, \hat{c}) \in \mathcal{T}$, c can determine $|\mathbf{f}'(e)|$.*

Proof. Denote by \mathcal{C} the connected components of $\mathcal{T} \setminus \mathcal{R}$. For $c \in \mathcal{T}$, denote by $C \in \mathcal{C}$ its connected component. We rewrite the (absolute value of) the multicommodity flow

$$|\mathbf{f}'(c, \hat{c})| = \sum_{\substack{c_1 \in \mathcal{T}_c \\ c_2 \notin \mathcal{T}_c}} \sum_{\{c_1, c_2\}_{uv} \in \mathcal{E}} \text{cap}(\{c_1, c_2\}_{uv}) = |\mathbf{f}'_1(c, \hat{c})| + |\mathbf{f}'_2(c, \hat{c})| - |\mathbf{f}'_3(c, \hat{c})|,$$

where

$$\begin{aligned}
|\mathbf{f}'_1(c, \hat{c})| &:= \sum_{\substack{c_1 \in \mathcal{T}_c \setminus C \\ c_2 \notin \mathcal{T}_c \setminus C}} \sum_{\{c_1, c_2\}_{uv} \in \mathcal{E}} \text{cap}(\{c_1, c_2\}_{uv}), \\
|\mathbf{f}'_2(c, \hat{c})| &:= \sum_{\substack{c_1 \in \mathcal{T}_c \cap C \\ c_2 \notin \mathcal{T}_c}} \sum_{\{c_1, c_2\}_{uv} \in \mathcal{E}} \text{cap}(\{c_1, c_2\}_{uv}), \text{ and} \\
|\mathbf{f}'_3(c, \hat{c})| &:= \sum_{\substack{c_1 \in \mathcal{T}_c \cap C \\ c_2 \in \mathcal{T}_c \setminus C}} \sum_{\{c_1, c_2\}_{uv} \in \mathcal{E}} \text{cap}(\{c_1, c_2\}_{uv}).
\end{aligned}$$

Note that $|\mathbf{f}'_1(c, \hat{c})|$ does only depend on the component of c , i.e., we need to determine and make known only $|\mathcal{C}| \in O(\sqrt{n})$ values to cover this term. For the other terms, we will reduce the problem to an aggregation on the spanning tree of C in the vain of [Lemma 8.1](#).

Concerning $|\mathbf{f}'_1(c, \hat{c})|$, we employ the following routine.

1. Using its spanning tree, each component $C \in \mathcal{C}$ determines a unique identifier (say, the smallest cluster identifier) and makes it known to all its nodes.
2. For each $\{c_1, c_2\}_{uv} \in \mathcal{E}$, u and v exchange their component identifiers.
3. The list of component identifiers and edges (C, \hat{C}) for each $(c, \hat{c}) \in \mathcal{R}$ is made known to all nodes. This enables each node to locally compute the tree resulting from contracting the components $C \in \mathcal{C}$ in \mathcal{T} .
4. For each $C \in \mathcal{C}$, fix an arbitrary $c \in C$. Each node $u \in \mathcal{T}_c \setminus C$ locally computes

$$\text{cap}_C(u) := \sum_{\substack{\{c_1, c_2\}_{uv} \in \mathcal{E} \\ c_2 \notin \mathcal{T}_c \setminus C}} \text{cap}(\{c_1, c_2\}_{uv});$$

we set $\text{cap}_c(u) := 0$ for all $u \notin \mathcal{T}_c \setminus C$ (nodes can determine whether they are in $\mathcal{T}_c \setminus C$ based on the information collected in the previous two steps).

5. For each $C \in \mathcal{C}$, make $\sum_{u \in V} \text{cap}_C(u)$ known to all nodes via a BFS tree of G . For all $c \in C$, we have that $|\mathbf{f}'_1(c, \hat{c})| = \sum_{u \in V} \text{cap}_C(u)$.

As discussed earlier, components' spanning trees have depth $\tilde{O}(\sqrt{n})$ and $|\mathcal{C}| \in O(\sqrt{n})$. Hence, Step 1 takes $\tilde{O}(\sqrt{n})$ rounds and Steps 3 and 5 take $O(\sqrt{n} + D)$ rounds. Step 2 requires only one round of communication and Step 4 is local. Overall, the routine requires $\tilde{O}(\sqrt{n} + D)$ rounds.

To determine $|\mathbf{f}'_2(c, \hat{c})|$ and $|\mathbf{f}'_3(c, \hat{c})|$ for each c , we proceed similarly to [Lemma 8.2](#).

1. For each $C \in \mathcal{C}$ and each $c \in C$, all nodes in c learn the list of ancestors of c that are in C (using the spanning tree of C in G).
2. For each $\{c_1, c_2\}_{uv} \in \mathcal{E}$, u and v exchange their component identifiers, as well as the ancestor lists determined in the previous step.
3. The list of component identifiers and edges (C, \hat{C}) for each $(c, \hat{c}) \in \mathcal{R}$ is made known to all nodes.
4. For each $C \in \mathcal{C}$, $c \in C$, and $u \in \mathcal{T}_c \cap C$, u locally computes

$$\text{cap}_c(u) := \sum_{\substack{\{c_1, c_2\}_{uv} \in \mathcal{E} \\ v \in \mathcal{T}_c \setminus C}} \text{cap}(\{c_1, c_2\}_{uv}) - \sum_{\substack{\{c_1, c_2\}_{uv} \in \mathcal{E} \\ v \in \mathcal{T}_c \setminus C}} \text{cap}(\{c_1, c_2\}_{uv}).^6$$

⁶If $v \in C$, u can decide whether $v \in \mathcal{T}_c$ based on the ancestor lists. If $v \notin C$, u can decide whether $v \in \mathcal{T}_c$ based on v 's component identifier and the information collected in Step 3.

5. For each edge $e = (c, \hat{c}) \in \mathcal{T}$, we aggregate $\sum_{u \in T_c \cap C} \text{cap}_c(u)$ on $T_c \cap C$, where T_c is the subtree of T (the spanning tree of G) corresponding to \mathcal{T}_c .

Note that, by definition of $\text{cap}_c(u)$, we have that

$$\sum_{u \in T_c \cap C} \text{cap}_c(u) = |\mathbf{f}'_2(c, \hat{c})| - |\mathbf{f}'_3(c, \hat{c})|.$$

Hence, it remains to analyze the running time of this second subroutine. Again, using that components' spanning trees have depth $\tilde{O}(\sqrt{n})$ and that $|C| \in O(\sqrt{n})$, we can conclude that Steps 1, 2, and 5 take $\tilde{O}(\sqrt{n})$ rounds, while Step 3 takes $O(\sqrt{n} + D)$ rounds. As Step 4 requires local computation only, the resulting running time is $\tilde{O}(\sqrt{n})$ rounds. Overall, we conclude that $|\mathbf{f}'(c, \hat{c})|$ can be computed for each c within $\tilde{O}(\sqrt{n} + D)$ rounds, by running each of the two subroutines and summing up their outputs. \square

8.2 Approximating \mathcal{G} by a Distribution over Simpler Graphs

Using the techniques of [26] and the above construction of low average stretch spanning trees, it is possible to design a distributed algorithm to compute a distribution of such spanning trees which approximates the cut structure of the underlying network graph within a factor $2^{O(\sqrt{\log n \log \log n})}$ (i.e., in the order of the average stretch of the computed spanning trees). However, when doing this, the number of spanning trees we need to compute can be linear in the size of \mathcal{G} . We follow the same general idea as Sherman, who applied the construction by Madry [19] recursively to decrease the *step* complexity, to avoid this sequential bottleneck and achieve a small *time* complexity in the distributed setting.

For each edge $e \in \mathcal{T}$, we define $\text{rload}_{\mathcal{T}}(e) := \text{cap}_{\mathcal{T}}(e)/\text{cap}(e) \geq 1$ to be the relative load of e (edges $e \in \mathcal{E} \setminus \mathcal{T}$ have $\text{rload}_{\mathcal{T}}(e) = 0$). The construction of [26] builds up a potential for each edge e of \mathcal{G} , where with each new tree added to the distribution, the potential of e grows by a term proportional to $\text{rload}_{\mathcal{T}}(e)/\max_{e' \in \mathcal{E}}\{\text{rload}_{\mathcal{T}}(e')\}$. The potential of each edge is bounded by $\alpha = n^{o(1)}$ and hence with every additional spanning tree, we are guaranteed to make progress for all edges $e \in \mathcal{E}$ with $\text{rload}_{\mathcal{T}}(e)$ close to $\max_{e' \in \mathcal{E}}\{\text{rload}_{\mathcal{T}}(e')\}$. In the worst case, this can just be a single edge for each spanning tree \mathcal{T} . The key idea of Madry [19] is to augment the tree \mathcal{T} with additional edges in order to reduce the maximum relative load so that in the new graph, a large number of edges have a relative load close to the maximum one.

Basically, we can reach a large number of edges with relative load close to the maximum relative load by repeatedly deleting the edge with largest relative load until a large number of the remaining edges has a relative load that is within a constant factor of the remaining maximum relative load. When deleting some edges of \mathcal{T} , one has to add back some of the original edges of \mathcal{G} in order to maintain the property that \mathcal{G} is embeddable into the resulting graph. Formally, let $\mathcal{F} \subseteq \mathcal{T}$ be a subset of the spanning tree edges. The edge set $\mathcal{T} \setminus \mathcal{F}$ defines a spanning forest of \mathcal{G} consisting of $|\mathcal{F}| + 1$ components. We define a subgraph $\mathcal{H}(\mathcal{T}, \mathcal{F})$ of \mathcal{G} as follows. The node set of $\mathcal{H}(\mathcal{T}, \mathcal{F})$ is \mathcal{V} . Further, $\mathcal{H}(\mathcal{T}, \mathcal{F})$ contains all edges in $\mathcal{T} \setminus \mathcal{F}$ and it contains all edges $\{c, c'\}_{uv} \in \mathcal{E}$ of \mathcal{G} for which c and c' are in different components in the forest induced by the edges in $\mathcal{T} \setminus \mathcal{F}$. Let $\mathcal{E}_{\mathcal{H}}$ be the set of edges of \mathcal{H} . We set the capacities $\text{cap}_{\mathcal{H}}(e)$ of edges $e \in \mathcal{E}_{\mathcal{H}}$ to be $\text{cap}_{\mathcal{H}}(e) := \text{cap}_{\mathcal{T}}(e)$ if $e \in \mathcal{T} \setminus \mathcal{F}$ and $\text{cap}_{\mathcal{H}}(e) := \text{cap}(e)$ otherwise. Note that this guarantees that \mathcal{G} is 1-embeddable into \mathcal{H} . For the following discussion, we define $\mathbb{H}[j]$ to be the set of graphs $\mathcal{H}(\mathcal{T}, \mathcal{F})$ for a spanning tree \mathcal{T} of \mathcal{G} and a set of edges \mathcal{F} of \mathcal{T} of size $|\mathcal{F}| \leq j$.

Assume that the weighted average stretch of the spanning tree \mathcal{T} as given by Eq. (2) is upper bounded by α . Also recall that we assume that all capacities of G are integers that are polynomially bounded in the number of nodes n . As throughout our construction, each edge capacity always approximately corresponds to the capacity of some cut in G , it is not hard to guarantee that all capacities of \mathcal{G} are integers between 1 and $\text{poly}(n)$. Given a spanning tree \mathcal{T} of \mathcal{G} , let $R := \max_{e \in \mathcal{T}}\{\text{rload}_{\mathcal{T}}(e)\}$ be the largest relative load of all edges of \mathcal{T} . In order to determine the set of edges \mathcal{F} , we start by partitioning the edges in \mathcal{T} into $i_{\max} = O(\log n)$ classes $\mathcal{F}_1, \dots, \mathcal{F}_{i_{\max}}$, where class \mathcal{F}_i contains all edges with relative load in $(R/2^i, R/2^{i-1}]$. Now, for any

$j_0 \leq |\mathcal{T}|$, there exists an edge class \mathcal{F}_i such that $|\bigcup_{i' < i} \mathcal{F}_{i'}| \leq j_0$ and $|\mathcal{F}_i| \geq j_0/i_{\max} = \Omega(j_0/\log n)$; otherwise, $|\mathcal{T}| = |\bigcup_i \mathcal{F}_i| < j_0 \leq |\mathcal{T}|$, a contradiction. We define $\mathcal{F}' := \bigcup_{i' < i} \mathcal{F}_{i'}$.

In [20], this set of edges is used to construct the graph $\mathcal{H}(\mathcal{T}, \mathcal{F}')$. For the distributed computation, it will be useful to have a graph $\mathcal{H}(\mathcal{T}, \mathcal{F})$ in which all the trees of the forest induced by $\mathcal{E}_{\mathcal{H}} \setminus \mathcal{F}$ have small diameter. We therefore rely on the same technique as for computing the capacities of \mathcal{T} and remove a few random additional edges of \mathcal{T} . In fact, we can simply use the same subset of edges $\mathcal{R} \subseteq \mathcal{T}$ that has been determined and used before, prior to Lemma 8.2. We define $\mathcal{F} := \mathcal{F}' \cup \mathcal{R}$ and use the graph $\mathcal{H} = (\mathcal{V}, \mathcal{E}_{\mathcal{H}}, \text{cap}_{\mathcal{H}}) := \mathcal{H}(\mathcal{T}, \mathcal{F})$. Since all the edges of \mathcal{T} with $\text{rload}_{\mathcal{T}}(E) > R/2^{i-1}$ are removed, all edges of \mathcal{H} have relative load at most $R/2^{i-1}$. Further, all the $\Omega(j/\log n)$ edges of \mathcal{F}_i have relative load larger than $R/2^i$. Based on Theorem 5.2 and Corollary 5.6 of [20] and on Theorem 3.1, we can show the following lemma.

Lemma 8.4. *Given are a distributed cluster (multi-)graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \text{cap}_{\mathcal{G}})$ consisting of $|\mathcal{V}| = N$ clusters and $|\mathcal{E}| = N \text{ polylog}(n)$ edges and a parameter $j \geq 1$ such that $j = \omega(\sqrt{n} \log n)$. There is a distributed algorithm to compute an $(2^{O(\sqrt{\log N \log \log N})}, \mathbb{H}[j])$ -distribution of \mathcal{G} on $2^{O(\sqrt{\log N \log \log N})} \cdot N/j$ graphs, which runs in the CONGEST model on the underlying network graph G in $(D + \sqrt{n}) \cdot n^{o(1)} \cdot N/j$ rounds.*

Proof. Let $\alpha = 2^{O(\sqrt{\log n \log \log n})}$ be the average stretch guarantee of the spanning tree algorithm. It follows directly from Theorem 5.2 and Lemma 5.5 in [20] that we can compute an $(O(\alpha), \mathbb{H}[|\mathcal{E}| \cdot \alpha \log(n)/s])$ -decomposition of \mathcal{G} on s graphs in time $s \cdot T_{\text{tree}}$ if the following conditions are satisfied:

- (1) The time for computing one low average stretch spanning tree is upper bounded by T_{tree} .
- (2) Given \mathcal{T} and the set of edges \mathcal{F} as computed above, let $\text{rload}_{\max} := \max_{e \in \mathcal{T} \setminus \mathcal{F}} \{\text{rload}_{\mathcal{T}}(e)\}$. The number of edges of $\mathcal{H}(\mathcal{T}, \mathcal{F})$ with relative load at least $\text{rload}_{\max}/2$ is $\Omega(|\mathcal{E}| \alpha / s)$.

As observed above, in \mathcal{T} , all edges in the set \mathcal{F}_i have relative load between $\text{rload}_{\max}/2$ and rload_{\max} . When constructing $\mathcal{H}(\mathcal{T}, \mathcal{F})$, the relative load of edges in $\mathcal{T} \setminus \mathcal{F}$ does not change and thus, all nodes in $\mathcal{F}_i \setminus \mathcal{F} = \mathcal{F}_i \setminus \mathcal{R}$ have relative load between $\text{rload}_{\max}/2$ and rload_{\max} . Recall that $|\mathcal{F}_i| = \Omega(j/\log n)$. By Lemma 8.2, with high probability, we have $|\mathcal{R}| = O(\sqrt{n})$. Since we assumed that $j = \omega(\sqrt{n} \log n)$, we have $|\mathcal{F}_i| = \omega(|\mathcal{R}|)$ and thus $|\mathcal{F}| = \Omega(j/\log n)$. The second condition is now satisfied by choosing $s = \Theta(|\mathcal{E}| \alpha \log(n)/j) = 2^{O(\sqrt{\log N \log \log N})} \cdot N/j$.

Assuming that the time to compute a single low average stretch spanning tree can be upper bounded by $T_{\text{tree}} = (D + \sqrt{n}) \cdot n^{o(1)}$, the lemma now follows. By Theorem 3.1, this is guaranteed as long as all edge lengths are integers between 1 and $2^{n^{o(1)}}$. Inspecting the construction in [20] and [26], we can observe that the edge lengths cannot get larger than a value exponential α . By rounding them to integers, we introduce an additional multiplicative error of factor 2, which does not affect the asymptotic behavior. As $\alpha = 2^{O(\sqrt{\log n \log \log n})}$, the claim of the lemma follows. \square

8.3 Transforming $\mathcal{H}(\mathcal{T}, \mathcal{F})$ into a j -Tree

Given a graph $\mathcal{H}(\mathcal{T}, \mathcal{F}) \in \mathbb{H}[j]$, it remains to transform $\mathcal{H}(\mathcal{T}, \mathcal{F})$ into an $O(j)$ -tree \mathcal{J} such that the two graphs are $O(1)$ -embeddable into each other. In the following, we first describe the construction and we formally prove that the resulting $O(j)$ -tree \mathcal{J} and the given graph $\mathcal{H}(\mathcal{T}, \mathcal{F})$ are $O(1)$ -embeddable into each other. We then show how to efficiently construct \mathcal{J} in a distributed way.

Assume that we are given a spanning tree \mathcal{T} of \mathcal{G} and a graph $\mathcal{H}(\mathcal{T}, \mathcal{F})$ which is constructed as described above. Consider the forest induced by the edges in $\mathcal{T} \setminus \mathcal{F}$.

Let $P_1 \subseteq \mathcal{V}$ be the set of clusters of $\mathcal{T} \setminus \mathcal{F}$ which are incident to one of the deleted tree edges in \mathcal{F} . We call P_1 the *primary portals* of $\mathcal{T} \setminus \mathcal{F}$. Given P_1 , we define the skeleton $\mathcal{S}_{\mathcal{T} \setminus \mathcal{F}}$ of $\mathcal{T} \setminus \mathcal{F}$ as follows. $\mathcal{S}_{\mathcal{T} \setminus \mathcal{F}}$ is obtained from \mathcal{T} by repeatedly deleting non-portal clusters of degree 1 until all remaining clusters

are either in P_1 or they have degree at least 2. Denote by P_2 all clusters of degree larger than 2 that are not primary portals; P_2 are the *secondary portals*. The set of all portal clusters now is $P := P_1 \cup P_2$. The skeleton $\mathcal{S}_{\mathcal{T} \setminus \mathcal{F}}$ is thus a forest consisting of a set of portals and paths connecting them, where all inner clusters of these paths have degree 2.

Given the skeleton $\mathcal{S}_{\mathcal{T} \setminus \mathcal{F}}$, consider one of these paths \mathcal{P} . In the last step, we remove the edge with the smallest capacity from each such \mathcal{P} . In doing so, we split the forest into trees so that each tree contains exactly one portal. It is straightforward to bound the number of resulting trees in terms of \mathcal{F} .

Lemma 8.5. *Let $\mathcal{H}(\mathcal{T}, \mathcal{F}) \in \mathbb{H}[j]$, i.e., $|\mathcal{F}| \leq j$. Then, in the above construction, the total number of portal nodes is less than $4j$.*

Proof. Clearly, $|P_1| \leq 2|\mathcal{F}| \leq 2j$. As when computing the skeleton $\mathcal{S}_{\mathcal{T} \setminus \mathcal{F}}$, non-portal clusters of degree 1 are successively removed, we obtain a forest whose leaves are primary portals. As the sum of the degrees in an N -node forest is at most $2(N - 1)$, the number of nodes of degree at least 3 is upper bounded by the number of leaves minus 2. We conclude that $|P_2| < |P_1| \leq 2j$, and hence $|P| < 4j$. \square

Finally, we identify each of the resulting trees with its portal and logically move all edges between different trees to the portals. For each edge $e = \{c, c'\} \in \mathcal{E}_{\mathcal{H}} \setminus (\mathcal{T} \setminus \mathcal{F})$ (i.e., each non-tree edge of $\mathcal{H}(\mathcal{T}, \mathcal{F})$), we add a virtual edge of capacity $\text{cap}_{\mathcal{G}}(e)$ between the portals of the trees containing c and c' , respectively. Further, let \mathcal{D} be the set of edges that were deleted from the paths of degree-2 clusters connecting portals in the skeleton. For every edge $e \in \mathcal{D}$, we add a virtual edge of capacity $\text{cap}_{\mathcal{H}}(e) = \text{cap}_{\mathcal{T}}(e)$ between the two portals that were connected by the path from which e was deleted.

Let us summarize this part of the construction; see [Figure 5](#) for an example of a possible result. Starting from a forest $\mathcal{T} \setminus \mathcal{F}$, do as follows:

1. Define P_1 as the endpoints of edges in \mathcal{F} ;
2. iteratively delete degree-1 clusters that are not in P_1 until this process halts;
3. define P_2 as the clusters retaining degree larger than 2 that are not in P_1 and set $P := P_1 \cup P_2$;
4. delete from each (maximal) path without clusters from P the edge $e \in \mathcal{D}$ of minimum capacity and replace it by an edge of the same capacity between its endpoints; and
5. for each edge $e \in \mathcal{E}_{\mathcal{H}}$ between different components of $\mathcal{T} \setminus (\mathcal{F} \cup \mathcal{D})$, add an edge of the same capacity between the unique portals in these components.

Hence, the resulting graph consists of the forest induced by $\mathcal{T} \setminus (\mathcal{F} \cup \mathcal{D})$ and (possibly parallel) edges between the unique portals of the trees of the forest. By [Lemma 8.5](#), the number of such portals is smaller than $4j$, implying that the resulting graph is a $4j$ -tree. In the following, we denote this $4j$ -tree by \mathcal{J} .

Mutual Embeddability of $\mathcal{H}(\mathcal{T}, \mathcal{F})$ and \mathcal{J} . Before discussing how to efficiently construct (and represent) \mathcal{J} in a distributed way, let us first show that $\mathcal{H}(\mathcal{T}, \mathcal{F})$ and \mathcal{J} are $O(1)$ -embeddable into each other.

The proofs of the following two lemmas is very similar to the corresponding result by Madry [\[19\]](#). However, since our j -tree construction slightly deviates from Madry's, the claims do not readily follow from any lemma in [\[19, 20\]](#).

Lemma 8.6. *$\mathcal{H}(\mathcal{T}, \mathcal{F})$ is $O(1)$ -embeddable into \mathcal{J} .*

Proof. There are three types of edges of $\mathcal{H}(\mathcal{T}, \mathcal{F})$ to distinguish:

- a) edges in $(\mathcal{T} \setminus \mathcal{F}) \setminus \mathcal{D}$,
- b) edges in \mathcal{D} , and
- c) the remaining edges from $\mathcal{E}_{\mathcal{H}}$ connecting different trees of the forest $\mathcal{T} \setminus \mathcal{F}$.

Case a) is the most straightforward, as all these edges are also present in \mathcal{J} with the same capacity. Edges from $e \in \mathcal{D}$ were deleted from a path \mathcal{P} connecting two portals in the skeleton. In \mathcal{J} , they can therefore be routed through the path \mathcal{P} and the virtual edge with capacity $\text{cap}_{\mathcal{T}}(e)$ connecting the portal nodes at the ends of \mathcal{P} . Because e is the lowest capacity edge of \mathcal{P} , this adds relative load at most 1 to each edge of \mathcal{P} .

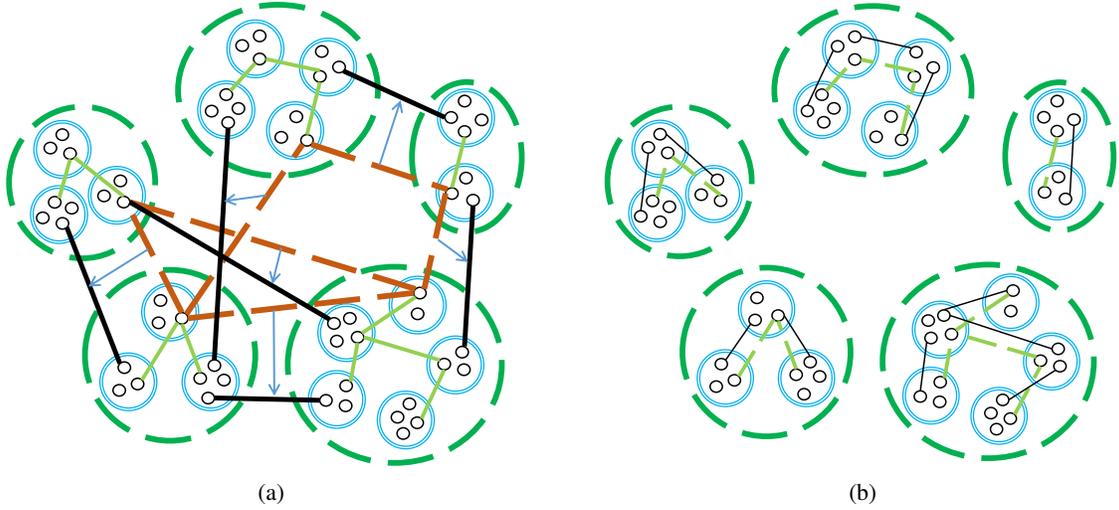


Figure 5: An example j -tree at the second level of recursion. On the left side, green circles indicate the components of the forest of this j -tree, which are each made of a number of 1-clusters, indicated by blue double-line circular shapes. Edges inside 1-clusters are not shown. Solid straight green edges indicate virtual edges of level-1 that became edges of the level-2 forest. For each of these green edges, there is a real edge between some two nodes of the same two level-1 components (not shown). Brown edges represent (virtual) core edges, and they are mapped (blue arrows) to corresponding real edges (solid black). On the right side, real edges related to the virtual forest edges are represented as solid black edges between two nodes of the same connected component.

Finally, let us consider one of the remaining edges $e \in \mathcal{E}_{\mathcal{H}}$. The edge $e = \{c_1, c_2\}$ connects two trees $T_1 \neq T_2$ of \mathcal{J} . Let us assume that $c_1 \in T_1$ and $c_2 \in T_2$. When routing from c_1 to c_2 , we follow

1. the path from c_1 in T_1 to the first skeleton cluster $s_1 \in T_1$ on the path to the (unique) portal $p_1 \in T_1 \cap P$,
2. the skeleton path from s_1 to p_1 ,
3. the virtual edge corresponding to e between p_1 and the (unique) portal $p_2 \in T_2 \cap P$,
4. the skeleton path from p_2 to the last skeleton cluster $s_2 \in T_2$ when going from p_2 to c_2 in T_2 , and
5. the path from s_2 to c_2 in T_2 .

Let us compare the path from c_1 via s_1 to p_1 with the path on which e is routed on the spanning tree \mathcal{T} . The part from c_1 to s_1 is also used when routing in \mathcal{T} . If from s_1 we follow the same direction to p_1 as in \mathcal{T} , the two paths are, in fact, identical up to the point when we reach p_1 . In this case, the capacities on this path suffice by construction, as we defined $\text{cap}_{\mathcal{T}}(e') = |\mathbf{f}'(e')|$. Let us therefore consider the case in which we set out in the opposite direction from s_1 on the skeleton path $\mathcal{P} \ni s_1$ connecting p_1 to some other portal p_2 than we would in \mathcal{T} . In that case, routing on \mathcal{T} would cross the edge $e' \in \mathcal{D}$ that was deleted from \mathcal{P} . Because e' is the edge from \mathcal{P} of smallest capacity, the contribution to the relative load of all edges crossed on \mathcal{P} is upper bounded by the relative load contributed to e' when routing on \mathcal{T} . Again, summing over all edges from $\mathcal{E}_{\mathcal{H}}$ falling under Case c), this may increase their total relative loads only by an additive 1. Trivially, the third step causes relative load 1 on the virtual edge corresponding to e , since it is not used for routing any other edge. Reasoning symmetrically for Steps 4 and 5, we can conclude that embedding $\mathcal{H}(\mathcal{T}, \mathcal{F})$ into \mathcal{J} leads to constant relative load on all edges. \square

Lemma 8.7. \mathcal{J} is $O(1)$ -embeddable into $\mathcal{H}(\mathcal{T}, \mathcal{F})$.

Proof. All the edges of the trees of \mathcal{J} are also present in $\mathcal{H}(\mathcal{T}, \mathcal{F})$ with the same capacity; they are hence straightforward to embed. Let us therefore consider the virtual edges connecting the portals of \mathcal{J} . There are two types of virtual edges, the ones representing edges from \mathcal{D} and those that correspond to the non-tree edges of $\mathcal{H}(\mathcal{T}, \mathcal{F})$. We first have a look at a virtual edge e corresponding to an edge $e' \in \mathcal{D}$. The edge e is

routed by following path from which e' was removed. Since the capacity of each edge on the path is at least the capacity of e , this contributes at most 1 to the relative load of each edge.

Now consider a virtual edge $e = \{c, c'\}$ corresponding to an edge $e' \in \mathcal{E}_{\mathcal{H}}$ of $\mathcal{H}(\mathcal{F}, \mathcal{T})$. The edge e is routed on the trees of \mathcal{J} the clusters c and c' reside in and via e' . The latter causes relative load 1, as e and e' have the same capacity and no other edge uses e' . Similarly to the embedding of $\mathcal{H}(\mathcal{T}, \mathcal{F})$ into \mathcal{J} , the tree parts of the routing path that are subpaths of the path between c and c' in \mathcal{T} and thus will not cause more than additive relative load 1 when summing over all edges of this type to embed. If we diverge from this path, this is because an edge from \mathcal{D} lies on the routing path in \mathcal{T} ; analogously to Lemma 8.6, following the skeleton path from which it was deleted to the respective portal increases the maximum relative load by at most an additional 1. \square

Distributed Implementation. Let us now move to the distributed implementation of the above $4j$ -tree construction. Recall that because \mathcal{F} includes the random set of edges \mathcal{R} , by Lemma 8.2, all trees in $\mathcal{T} \setminus \mathcal{F}$ have depth $\tilde{O}(\sqrt{n})$. With this in mind, constructing the skeleton is fairly simple.

Lemma 8.8. *Given are a spanning tree \mathcal{T} of a distributed cluster graph and the set of tree edges \mathcal{F} as computed above. We can determine the skeleton $\mathcal{S}_{\mathcal{T} \setminus \mathcal{F}}$, the set of portals P , and the set of edges \mathcal{D} (i.e., for $\{c, c'\}_{uv} \in \mathcal{D}$, u and v will learn this) in time $O(\sqrt{n} \log n)$ in the CONGEST model on the underlying network graph. In the same time, we can also orient the trees rooted at the portals.*

Proof. W.l.o.g., consider a single tree T of the forest $\mathcal{T} \setminus \mathcal{F}$. By Lemma 8.2, the induced tree in G has depth $\tilde{O}(\sqrt{n})$. Perform the following steps:

- For each edge $e \in \mathcal{F}$, its incident clusters learn⁷ that they are primary portals, i.e., are in P_1 .
- Iteratively mark non-portal clusters with at most 1 marked neighboring cluster until this process stops. Unmarked clusters are in the skeleton.
- Unmarked clusters with more than two unmarked neighboring portals are secondary portals.
- The skeleton paths connecting portals find a minimum capacity edge and add it to \mathcal{D} .
- Each tree of $\mathcal{T} \setminus (\mathcal{F} \cup \mathcal{D})$ is rooted at its unique portal, whose identifier is made known to all nodes in the induced tree in G (together with clusters' spanning trees).
- These identifiers are exchanged with all neighbors in G .

From the gathered information, for each edge $\{c, c'\}_{uv} \in \mathcal{J}$, u and v now can determine its membership and its capacity in \mathcal{J} . Observe that the bound of $\tilde{O}(\sqrt{n})$ on the depth of the spanning trees of G leveraged for communication in the above construction implies that all the above steps can be completed in $\tilde{O}(\sqrt{n})$ rounds, which completes the proof. \square

The trees rooted at the portals now induce the clusters of the new cluster graph.

Corollary 8.9. *Given a graph $\mathcal{H}(\mathcal{T}, \mathcal{F}) \in \mathbb{H}[j/4]$ as computed above on a cluster graph whose clusters' spanning trees have maximum depth d , there is an $\tilde{O}(D + d + \sqrt{n})$ -round distributed algorithm to compute*

- a cluster graph whose clusters' spanning trees have depth $d + \tilde{O}(\sqrt{n})$; and
- a j -tree \mathcal{J} on this cluster graph, i.e., for each edge $e \in \mathcal{J}$, there is a corresponding graph edge $\{u, v\} \in E$ whose constituent nodes know that $e \in \mathcal{J}$ as well as $\text{cap}_{\mathcal{J}}(e)$; such that
- $\mathcal{H}(\mathcal{T}, \mathcal{F})$ is 1-embeddable into \mathcal{J} and \mathcal{J} is $O(1)$ -embeddable into $\mathcal{H}(\mathcal{T}, \mathcal{F})$; and
- the new clusters are induced by the tree components of \mathcal{J} .

Proof. This readily follows from Lemmas 8.2, 8.5, 8.6, 8.7, and 8.8. The only thing left to note is that clusters can learn the number of nodes they contain by a simple converge- and broadcast operation on their spanning trees. \square

⁷A cluster for which edges to children are in \mathcal{F} may not “know” about its incident edges in \mathcal{F} as a whole, but determining whether there is at least one is trivial.

8.4 Sampling from the Recursively Constructed Distribution

We have now all pieces in place to efficiently sample from a distribution similar to Sherman's in a distributed fashion. The difference is that [Theorem 3.1](#) and thus [Lemma 8.4](#) merely give $\alpha \in 2^{O(\sqrt{\log n \log \log n})}$, implying that we must use fewer levels of recursion to ensure that the final approximation guarantee of the congestion approximator will remain in $n^{o(1)}$.

Theorem 8.10. *W.h.p., within $(\sqrt{n} + D)n^{o(1)}$ rounds of the CONGEST model, we can sample a tree \mathcal{T} from a distribution of $n^{1+o(1)}$ (virtual) rooted spanning trees on G with the following properties.*

- For any cut of G of capacity C , the capacity of the cut in \mathcal{T} is at least C .
- For any cut of G of capacity C , the expected capacity of the cut in \mathcal{T} is at most αC , where $\alpha \in n^{o(1)}$.
- The distributed representation of \mathcal{T} is given by a hierarchy of cluster graphs $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i, \mathcal{L}_i, \mathcal{T}_i, \psi_i)$, $i \in \{0, \dots, i_0\}$, $i_0 \in o(\log n)$, on network graph G , with the following properties.
 - The spanning trees of the clusters of \mathcal{G}_i have depth $\tilde{O}(\sqrt{n})$.
 - $|\mathcal{V}_{i_0}| = n^{1/2+o(1)}$.
 - \mathcal{G}_i is the (rooted) tree resulting from \mathcal{T} by contracting the clusters of \mathcal{G}_i .
 - For $i > 0$, \mathcal{G}_i is also a cluster graph on network graph \mathcal{G}_{i-1} .
 - For $i > 0$, each cluster $c_i \in \mathcal{V}_i$ of \mathcal{G}_i , interpreted as cluster graph on \mathcal{G}_{i-1} , contains a unique portal cluster $p(c_i) \in \mathcal{V}_{i-1}$ of \mathcal{G}_{i-1} that is incident⁸ to all edges of \mathcal{G}_i containing c_i . That is, \mathcal{G}_{i-1} is a $|\mathcal{V}_i|$ -tree with core $p(\mathcal{V}_i)$.

Proof. In the following, we will use w.h.p. statements as if they were deterministic; the result then follows by taking the union bound over all (polynomially many in n) such statements we use.

Set $\beta := 2^{\log^{3/4} n}$. To start the recursion, we will use G as cluster graph of itself. Formally, $\tilde{\mathcal{G}}_0 := (V, E, V, \{\{\{v\}, \emptyset\}_{v \in V}, \text{id}\})$, where id is the identity function. We perform the following construction until it terminates:

1. Sparsify $\tilde{\mathcal{G}}_{i-1}$ using [Lemma 6.1](#) for some fixed constant ε , e.g., $\varepsilon = 1/2$. This takes $(\sqrt{n} + D)n^{o(1)}$ rounds. Multiply all edge capacities by $1/(1 - \varepsilon)$ (so $\tilde{\mathcal{G}}_{i-1}$ can be 1-embedded into the sparser graph).
2. If $|\mathcal{V}_{i-1}| \notin \omega(\sqrt{n}\beta/\log n)$, set $i_0 := i$ and stop. This takes $O(D)$ rounds by communicating over a BFS tree of G .
3. Apply [Lemma 8.4](#) for $j = |\mathcal{V}_{i-1}|/(4\beta)$ to the sparsified cluster graph; by the previous step, this choice of j is feasible. As $|\mathcal{V}_{i-1}|/j = 4\beta \in n^{o(1)}$, constructing the distribution requires $(\sqrt{n} + D)n^{o(1)}$ rounds in total.
4. Sample a cluster graph from the distribution. This is done in $O(D)$ rounds letting some node broadcast $O(\log n)$ random bits over a BFS tree.
5. Apply [Corollary 8.9](#) to extract a $|\mathcal{V}_{i-1}|/\beta$ -tree of \mathcal{G}_{i-1} . The corollary also yields a cluster graph $\tilde{\mathcal{G}}_i$ (which is also a cluster graph on network graph \mathcal{G}_{i-1}) so that each of its clusters c_i contains exactly one portal cluster $p(c_i)$ of the $|\mathcal{V}_{i-1}|/\beta$ -tree on \mathcal{G}_{i-1} . This step completes in $\tilde{O}(\sqrt{n} + D)$ rounds: there are fewer than $\log_\beta \sqrt{n} \ll \log n$ iterations of the overall construction, as $|\mathcal{V}_i| \leq |\mathcal{V}_{i-1}|/\beta$, implying that $d \in \tilde{O}(\sqrt{n})$ for each application of [Corollary 8.9](#).
6. Recurse on $\tilde{\mathcal{G}}_i$, i.e., set $i := i + 1$ and go back to Step 1.

When the above construction halts, we have that $|\mathcal{V}_{i_0-1}| = O(\sqrt{n}\beta) = n^{1/2+o(1)}$. Thus, we can make the (sparsified) cluster graph $|\mathcal{G}_{i_0-1}|$ known to all nodes in $(\sqrt{n} + D)n^{o(1)}$ rounds via a BFS tree of G . We then continue the construction locally without controlling the size of components, which removes the constraint on j when applying [Lemma 8.8](#), until the core becomes empty, i.e., we construct a tree.⁹ We collapse the cluster graph hierarchy for all locally performed iterations $i \geq i_0$, which defines the tree \mathcal{G}_{i_0} on clusters \mathcal{C}_{i_0} (this is feasible as each \mathcal{G}_i , $i > 0$, is also a cluster graph on network graph \mathcal{G}_{i-1}).

⁸Note that the corresponding physical edges in G may still connect to different sub-clusters of c_i .

⁹This is essentially Sherman's construction on the small constructed cluster graph.

This completes the description of the algorithm. Summing up the running times of the individual steps and using that $i_0 = o(\log n)$, we conclude that the construction takes $(\sqrt{n} + D)n^{o(1)}$ rounds. The construction also maintained the stated structural properties of the cluster hierarchy. Hence, it remains to show that (i) we sampled from a distribution of $n^{1+o(1)}$ trees and (ii) the stated cut approximation properties are satisfied.

Showing these properties now is straightforward. In each step $i > 0$ of the recursion, by [Lemma 8.4](#) we constructed a distribution on $\tilde{O}(\beta) |\mathcal{V}_{i-1}|$ -trees. The total number of recursive steps (including the local ones), is bounded by $\lceil \log_\beta n \rceil = O(\log^{1/4} n)$, as $|\mathcal{V}_i| \leq |\mathcal{V}_{i-1}|/\beta$ for each $i > 0$. On each level of recursion, we compute a distribution on $2^{O(\sqrt{\log |\mathcal{V}_i| \log \log |\mathcal{V}_i|})} \beta \leq 2^{O(\sqrt{\log n \log \log n})} \beta$ graphs. Hence, the total number of virtual trees in the (implicit) distribution of virtual trees from which we sampled is bounded by

$$\left(2^{O(\sqrt{\log n \log \log n})} \beta\right)^{\lceil \log_\beta n \rceil} = n \cdot 2^{O(\sqrt{\log n \log \log n} \log^{1/4} n)} = n^{1+o(1)}.$$

Consider a cut of G of capacity C . By the properties of decompositions and the fact that we multiplied capacities by $1/(1-\varepsilon)$ whenever we sparsified, G is 1-embeddable into any of the trees we might construct, implying that the corresponding cut of the sampled tree has capacity at least C . As in each step, we (i) apply a $(1+\varepsilon)$ -sparsifier and multiply capacities by $1/(1-\varepsilon)$ for constant ε , (ii) construct a $(2^{O(\sqrt{\log n \log \log n})}, \mathbb{H})$ -decomposition (for some family \mathbb{H}) from which we sample, and (iii) transform the resulting graph into a j -tree which can be $O(1)$ -embedded into the graph from which it is constructed, we overestimate the capacity of a given cut by an expected factor of $2^{O(\sqrt{\log n \log \log n})} \cdot O(1) = 2^{O(\sqrt{\log n \log \log n})}$ in each step. Using that this bound is uniform and the randomness on each level of recursion is independent, it follows that the expected capacity of a cut of G of capacity C in the sampled virtual tree is bounded by

$$\left(2^{O(\sqrt{\log n \log \log n})}\right)^{\lceil \log_\beta n \rceil} = 2^{O(\sqrt{\log n \log \log n} \log^{1/4} n)} = n^{o(1)}.$$

□

9 The High-Level Algorithm

The algorithm is a distributed implementation of Sherman's algorithm [29]. It consists of a logarithmic number of calls to algorithm `AlmostRoute`, described in [Section 9.1](#), and one computation of a maximum-weight spanning tree and routing the left-over demand through this tree. Pseudocode for the top-level algorithm is presented in [Algorithm 1](#).

Algorithm 1 Max Flow. Input: demand vector $\mathbf{b} \in \mathbb{R}^n$; output: flow vector $\mathbf{f} \in \mathbb{R}^m$.

- 1: $\mathbf{b}_0 \leftarrow \mathbf{b}; \mathbf{f}_0 \leftarrow \mathbf{0}$
 - 2: **for** $i \leftarrow 1$ to $(\log m + 1)$ **do**
 - 3: $\mathbf{f}_i \leftarrow \text{AlmostRoute}(\mathbf{b}_i, \frac{1}{2})$
 - 4: $\mathbf{b}_i \leftarrow \mathbf{b}_{i-1} - B\mathbf{f}_{i-1}$.
 - 5: Compute a maximum weight spanning tree T on G , where weights are the capacities of edges.
 - 6: Route the residual demand \mathbf{b}_i through T ; let \mathbf{f}_T be the resulting flow.
 - 7: Output $\mathbf{f}_T + \sum_{i=1}^{1+\log m} \mathbf{f}_i$.
-

Most of this section is dedicated to explaining how to implement the `AlmostRoute` algorithm. Let us first quickly outline how we implement the final steps using standard techniques.

Lemma 9.1. *Steps 5–6 Can be implemented in the CONGEST model in $\tilde{O}(D + \sqrt{n})$ rounds w.h.p.*

Proof sketch. A maximum weight spanning tree T can be computed in $\tilde{O}(D + \sqrt{n})$ rounds using the minimum weight spanning tree algorithm of Kuttan and Peleg [18] (say, by assigning weight $w(e) := -\text{cap}(e)$ for each edge e). To compute the flow, we use the following observation: if T was rooted at one of its nodes, then to route the demand \mathbf{b}_t over T , it would be sufficient for each node v to learn the total demand d_v in the subtree rooted at v . In this case each node v assigns d_v units of flow to the edge leading from v to its parent.

We now show how to root the tree and find the total demand in each subtree in $\tilde{O}(D + \sqrt{n})$ rounds. The algorithm is as follows. Remove each edge of the tree independently with probability $1/\sqrt{n}$. W.h.p.,

- (i) each connected component induced by the remaining edges contains has strong diameter $\tilde{O}(\sqrt{n})$,
- (ii) $O(\sqrt{n})$ edges are removed, and hence
- (iii) the number of components is $O(\sqrt{n})$.

Within each component, all demands are summed up, and this sum is made known to all nodes. The summation takes $\tilde{O}(\sqrt{n})$ rounds due to (i), and we can pipeline the announcement of the sums over a BFS tree in $\tilde{O}(\sqrt{n} + D)$ rounds due to (iii).

Moreover, in this time we can also assign unique identifiers to the components (e.g. the minimum identifier) and make the tree resulting from contracting components globally known. Using local computation only, nodes then can root this tree (e.g. at the cluster of minimum identifier) and determine the sum the demands of the clusters that are fully contained in their subtree. Using a simple broadcast, the orientation of edges within components is determined, and using a convergecast on the components, each node can determine the sum of demands in its subtree. These steps take another $\tilde{O}(\sqrt{n})$ rounds. \square

9.1 Algorithm AlmostRoute: The Gradient Descent

We now explain how to implement Algorithm AlmostRoute in a distributed setting. The idea is to use gradient descent with the potential function

$$\phi(\mathbf{f}) = \text{smax}(C^{-1}\mathbf{f}) + \text{smax}(2\alpha R(\mathbf{b} - B\mathbf{f})),$$

where the “soft-max” function, defined by

$$\text{smax}(\mathbf{y}) = \log \left(\sum_{i=1}^k e^{y_i} + e^{-y_i} \right) \quad \text{for all } \mathbf{y} \in \mathbb{R}^k,$$

is used as a differentiable approximation to the “max” function.

Given this potential function, AlmostRoute performs $O(\alpha^2 \varepsilon^{-3} \log n)$ updates on \mathbf{f} and outputs a flow \mathbf{f} that optimizes the potential function up to a $(1 + \varepsilon)$ factor.¹⁰ Pseudocode for this algorithm is given in Algorithm 2.

¹⁰Sherman claims that one can save a factor of $1/\varepsilon$ by a more careful scaling [29].

Algorithm 2 AlmostRoute(\mathbf{b}, ε)

```
1:  $k_b \leftarrow 2\alpha \|R\mathbf{b}\|_\infty \varepsilon / (16 \log n)$ ;  $\mathbf{b} \leftarrow k_b \mathbf{b}$ .
2: repeat
3:    $k_f \leftarrow 1$ 
4:   while  $\phi(\mathbf{f}) < 16\varepsilon^{-1} \log n$  do
5:      $\mathbf{f} \leftarrow \mathbf{f} \cdot (17/16)$ ;  $\mathbf{b} \leftarrow \mathbf{b} \cdot (17/16)$ ;  $k_f \leftarrow k_f \cdot (17/16)$ 
6:      $\delta \leftarrow \sum_{e \in E} |\text{cap}(e) \frac{\partial \phi}{\partial f_e}|$ 
7:     if  $\delta \geq \varepsilon/4$  then
8:        $f_e \leftarrow f_e - \text{sgn}\left(\frac{\partial \phi}{\partial f_e}\right) \cdot \text{cap}(e) \frac{\delta}{1+4\alpha^2}$ 
9:     else
10:       $f_e \leftarrow f_e/k_f$  for all edges  $e \in E$ .
11:       $b_v \leftarrow b_v/(k_b k_f)$  for all nodes  $v \in V$ 
12:    return
13: until done
```

To implement this algorithm in a distributed setting, we need to compute R , and to do multiplications by R or its transpose R^\top , distributively. These multiplications are needed for computing $\phi(\mathbf{f})$ and its partial derivatives. We remark that R and R^\top are not constructed explicitly, as we need to ensure a small time complexity for each iteration. Assuming that we can perform these operations, each step of AlmostRoute can be completed in $\tilde{O}(D)$ additional rounds.

We maintain the invariant that at the beginning of each iteration of the **repeat** loop, each node v knows the current flow over each of the links v is incident to, and the current demand at v (i.e., $(\mathbf{b} - B\mathbf{f})_v$). Let us break the potential function ϕ in two, i.e.,

$$\phi(\mathbf{f}) = \phi_1(\mathbf{f}) + \phi_2(\mathbf{f}), \quad \text{where } \phi_1(\mathbf{f}) = \text{smax}(C^{-1}\mathbf{f}) \quad \text{and} \quad \phi_2(\mathbf{f}) = \text{smax}(2\alpha R(\mathbf{b} - B\mathbf{f})).$$

We proceed as follows. First, we compute $\phi_1(\mathbf{f})$: to find $\text{smax}(C^{-1}\mathbf{f})$, it suffices to sum $\exp(f_e/\text{cap}(e))$ and $\exp(-f_e/\text{cap}(e))$ over all edges e , which can be done in $O(D)$ rounds. As Sherman points out, $\phi(\mathbf{f}) = \Theta(\varepsilon^{-1} \log n)$ due to the scaling, and thus, encoding $\exp(\phi(\mathbf{f}))$ with sufficient accuracy requires $O(\varepsilon^{-1} \log n)$ bits, which is thereby also a bound on the encoding length of all individual terms in the sums for ϕ_1 and ϕ_2 . The error introduced by rounding these values to integers is small enough to not affect the asymptotics of the running time.

For determining $\phi_2(\mathbf{f})$, we compute the vector $\mathbf{y} := 2\alpha R(\mathbf{b} - B\mathbf{f})$ and then do an aggregation on a BFS tree as for $\phi_1(\mathbf{f})$. Since $B\mathbf{f}$ can be computed instantly ($(B\mathbf{f})_v$ is exactly the net flow into v), this boils down to multiplying a locally known vector with R . Before we discuss how to implement this operation, let us explain more about the structure of R and how we determine $\frac{\partial \phi}{\partial f_e}$, which is required in Lines 6 and 8 of the algorithm.

The linear operator R is induced by graph cuts. More precisely, in the matrix representation of R , there is one row for each cut our congestion approximator (explicitly) considers. We will clarify the structure of R shortly; for now, denote by I the set of row indices of R . Observe that

$$\frac{\partial \phi}{\partial f_e} = \frac{\exp(f_e/\text{cap}(e)) - \exp(-f_e/\text{cap}(e))}{\text{cap}(e) \exp(\phi_1)} + \frac{\partial \phi_2}{\partial f_e} \quad (3)$$

and hence, given that ϕ_1 is known, the first term is locally computable. The second term expands to

$$\frac{\partial \phi_2}{\partial f_e} = \sum_{i \in I} \frac{\partial \phi_2}{\partial y_i} \cdot \frac{\partial y_i}{\partial f_e} = \sum_{i \in I} \frac{\exp(y_i) - \exp(-y_i)}{\exp(\phi_2)} \cdot \frac{2\alpha B_{i,e}}{\text{cap}(i)},$$

where $\text{cap}(i)$ is the capacity of cut i in the congestion approximator and $B_{i,e} \in \{-1, 0, 1\}$ denotes whether e is outgoing (-1), ingoing (1), or not crossing cut i .¹¹

The cuts $i \in I$ are induced by the edges of a collection of (rooted, virtual, capacitated) spanning trees \mathbb{T} , where for $\mathcal{T} \in \mathbb{T}$ we write $(v, \hat{v}) \in \mathcal{T}$ if \hat{v} is the parent of v and denote by \mathcal{T}_v the subtree rooted at v . For each $\mathcal{T} \in \mathbb{T}$, each edge $(v, \hat{v}) \in \mathcal{T}$ now induces a (directed) cut $(T_v; \overline{T}_v)$ with index $i(\mathcal{T}, (v, \hat{v}))$. We denote the set of edges crossing this cut by $\delta(\mathcal{T}_v)$. Let us also define

$$p(\mathcal{T}, v) = \frac{\exp(y_{i(\mathcal{T}, (v, \hat{v}))}) - \exp(-y_{i(\mathcal{T}, (v, \hat{v}))})}{\exp(\phi_2)} \cdot \frac{2\alpha}{\text{cap}_{\mathcal{T}}((v, \hat{v}))}.$$

With this notation, we have that

$$\frac{\partial \phi_2}{\partial f_e} = \sum_{\mathcal{T} \in \mathbb{T}} \sum_{\substack{(v, \hat{v}) \in \mathcal{T} \\ e \in \delta(\mathcal{T}_v)}} p(\mathcal{T}, v) \cdot B_{i(\mathcal{T}, (v, \hat{v})), e}.$$

We call $p(\mathcal{T}, (v, \hat{v}))$ the *price* of the (virtual) edge $(v, \hat{v}) \in \mathcal{T}$. Let $\mathcal{P}_{v, \mathcal{T}}$ denote the edge set of the unique path in \mathcal{T} from v to the root of \mathcal{T} . We define a *node potential* for each node v by

$$\pi_v := \sum_{\mathcal{T} \in \mathbb{T}} \sum_{(w, \hat{w}) \in \mathcal{P}_{v, \mathcal{T}}} p(\mathcal{T}, (w, \hat{w})).$$

For any $e = (u, v)$, the cuts induced by edges in $\mathcal{T} \in \mathbb{T}$ that e crosses correspond to the edges on the unique path from u to v in \mathcal{T} . For all edges $(w, \hat{w}) \in \mathcal{T}$ on the path from u to the least common ancestor of u and v in \mathcal{T} , $B_{i(\mathcal{T}, (w, \hat{w})), e} = -1$, while $B_{i(\mathcal{T}, (w, \hat{w})), e} = +1$ for the edges on the path between v and this least common ancestor. Thus,

$$\frac{\partial \phi_2}{\partial f_e} = \pi_v - \pi_u, \quad (4)$$

and our task boils down to determining the value of the potential π_v at each node $v \in V$. To this end, we need two key subroutines to compute distributively the following quantities.

- (1) y_i for each cut i . Note that $\mathbf{b} - B\mathbf{f}$ is known distributedly, i.e., each node knows its own coordinate of this vector. For each tree in $\mathcal{T} \in \mathbb{T}$, we need to aggregate this information from the leaves to the root. This means to simulate a convergecast on the virtual tree \mathcal{T} .
- (2) π_v for each node v . Provided that each (virtual) tree edge knows its y -value and ϕ_2 , the prices can be computed locally. Then the contribution of each tree to the node potentials can be computed by a downcast from the corresponding root to its leaves.

With these routines, one iteration of the **repeat** loop is now executed as follows:

1. Compute ϕ_1 , \mathbf{y} (local knowledge), and ϕ_2 (aggregation on BFS tree once \mathbf{y} is known).
2. Check the condition in Line 4. If it holds, locally update \mathbf{b} , \mathbf{f} , and k_f , and go to the previous step.
3. Compute the potential π (local knowledge).
4. For each $e \in E$, its incident edges determine $\frac{\partial \phi}{\partial f_e}$ (based on Equations 3 and 4, it suffices to exchange π_u and π_v over e).
5. Compute δ (aggregation on BFS tree).
6. Locally update f_e and b_v for all $e \in E$ and $v \in V$.

Note that all of the individual operations except for computation of \mathbf{y} and π can be completed in $O(D)$ rounds. Sherman proved [29] that AlmostRoute terminates after $\tilde{O}(\varepsilon^{-3}\alpha^2)$ iterations. As it is only called $O(\log n)$ times by the max-flow algorithm, Theorem 1.1 follows if we can compute \mathbf{y} and π in $(\sqrt{n} + D)n^{o(1)}$ rounds for an α -congestion approximator with $\alpha = n^{o(1)}$; this is subject of the next subsection.

¹¹Technically, $B_{i,e} = \sum_{v \in S_i} B_{ve}$ where S_i is the set of nodes defining cut i .

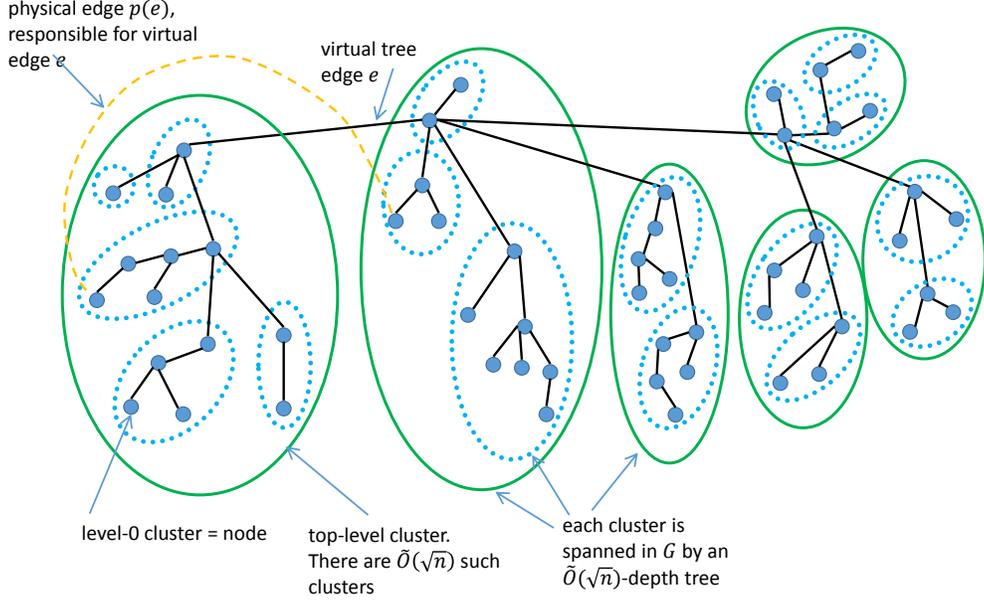


Figure 6: Hierarchical cluster decomposition of a virtual tree $\mathcal{T} \in \mathbb{T}$. Black edges are virtual tree edges, which are represented by a physical edge connecting the top-level clusters they connect (the orange dotted edge $p(e)$ corresponds to the edge labeled e). Each cluster is spanned by a tree in G of depth $\tilde{O}(\sqrt{n})$, which is not shown.

9.2 Congestion Approximation

Our congestion approximator R is defined by the edge-induced cuts of a sample \mathbb{T} of virtual trees \mathcal{T} from a recursively constructed distribution. The trees are represented distributedly by a hierarchy of *cluster graphs* (see Figure 6 for an illustration and Section 5 for the formal definition of cluster graphs). Intuitively, a cluster graph partitions the nodes into clusters, each of which has a spanning tree rooted at a leader, and a collection of edges between clusters that are represented by corresponding graph edges between some nodes of the clusters they connect. In Section 8, we have shown the following theorem.

Theorem 8.10. (restated) *W.h.p., within $(\sqrt{n} + D)n^{o(1)}$ rounds of the CONGEST model, we can sample a tree \mathcal{T} from a distribution of $n^{1+o(1)}$ (virtual) rooted spanning trees on G with the following properties.*

- For any cut of G of capacity C , the capacity of the cut in \mathcal{T} is at least C .
- For any cut of G of capacity C , the expected capacity of the cut in \mathcal{T} is at most αC , where $\alpha \in n^{o(1)}$.
- The distributed representation of \mathcal{T} is given by a hierarchy of cluster graphs $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i, \mathcal{L}_i, \mathcal{T}_i, \psi_i)$, $i \in \{0, \dots, i_0\}$, $i_0 \in o(\log n)$, on network graph G , with the following properties.
 - The spanning trees of the clusters of \mathcal{G}_i have depth $\tilde{O}(\sqrt{n})$.
 - $|\mathcal{V}_{i_0}| = n^{1/2+o(1)}$.
 - \mathcal{G}_i is the (rooted) tree resulting from \mathcal{T} by contracting the clusters of \mathcal{G}_i .
 - For $i > 0$, \mathcal{G}_i is also a cluster graph on network graph \mathcal{G}_{i-1} .
 - For $i > 0$, each cluster $c_i \in \mathcal{V}_i$ of \mathcal{G}_i , interpreted as cluster graph on \mathcal{G}_{i-1} , contains a unique portal cluster $p(c_i) \in \mathcal{V}_{i-1}$ of \mathcal{G}_{i-1} that is incident¹² to all edges of \mathcal{G}_i containing c_i . That is, \mathcal{G}_{i-1} is a $|\mathcal{V}_i|$ -tree with core $p(\mathcal{V}_i)$.

The first two properties of each \mathcal{T} stated in the theorem imply that we can use them to construct a good congestion approximator R . More precisely, Lemma 3.3 implies the following corollary.

¹²Note that the corresponding physical edges in G may still connect to different sub-clusters of c_i .

Corollary 9.2. *Sampling a collection \mathbb{T} of $O(\log n)$ virtual trees given by [Theorem 3.2](#) and using them as congestion approximator R in the way specified in [Section 9.1](#) implies that the total number of iterations of [Algorithm 2](#) is $n^{o(1)}$.*

All that remains now is to show that the distributed representation of each sampled $\mathcal{T} \in \mathbb{T}$ allows to simulate a convergecast and a downcast on \mathcal{T} in $(\sqrt{n} + D)n^{o(1)}$ rounds: then we can implement the key subroutines (1) and (2) (i.e., compute \mathbf{y} and $\boldsymbol{\pi}$) outlined in [Section 9.1](#) with this time complexity, and by [Corollary 9.2](#) the total number of rounds of the computation is bounded by $(\sqrt{n} + D)n^{o(1)}$.

Fortunately, the recursive structure of the decomposition is very specific. The cluster graphs of the different levels of recursion are nested, i.e., the clusters of the $(i - 1)^{\text{th}}$ level of recursion are subdivisions of the clusters of the i^{th} level. What is more, each cluster is a subtree of the virtual tree and is spanned by a tree of depth $\tilde{O}(\sqrt{n})$ in G (cf. [Figure 6](#)) Hence, while the physical graph edges representing the virtual tree edges are between arbitrary nodes within the clusters they connect, we can (i) identify each cluster on each hierarchy level with the root of the subtree induced by its nodes, (ii) handle such subtrees recursively (both for convergecasts and downcasts), (iii) on each level of recursion but the last, perform the relevant communication by broadcasting or upcasting on the underlying cluster spanning trees in G of depth $\tilde{O}(\sqrt{n})$, and (iv) communicate over a BFS tree of G on the final level of recursion, where merely $n^{1/2+o(1)}$ clusters/nodes of the virtual tree remain.

Corollary 9.3. *On each virtual tree $\mathcal{T} \in \mathbb{T}$, we can simulate convergecast and upcast operations in $\tilde{O}(\sqrt{n} + D)$ rounds.*

[Theorem 1.1](#) now follows from Sherman’s results on the number of iterations of the gradient descent algorithm [[29](#)], the discussion in [Section 9.1](#), and [Corollaries 9.2](#) and [9.3](#).

References

- [1] I. Abraham, Y. Bartal, and O. Neiman. Nearly tight low stretch spanning trees. In *Proc. of the Symp. on Found. of Comp. Sci. (FOCS)*, pages 781–790. IEEE, 2008.
- [2] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows*. Prentice-Hall, Engelwood Cliffs, New Jersey, 1993.
- [3] N. Alon, R. M. Karp, D. Peleg, and D. West. A graph-theoretic game and its application to the k -server problem. *SIAM Journal on Computing*, 24(1):78–100, 1995.
- [4] N. Alon, R. M. Karp, D. Peleg, and D. West. A graph-theoretic game and its application to the k -server problem. *SIAM J. on Computing*, 24(1):78–100, 1995.
- [5] S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.
- [6] B. Awerbuch. Reducing complexities of the distributed max-flow and breadth-first-search algorithms by means of network synchronization. *Networks*, 15(4):425–437, Winter 1985.
- [7] B. Awerbuch and R. Khandekar. Stateless distributed gradient descent for positive linear programs. *SIAM Journal on Computing*, 38(6):2468–2486, 2009.
- [8] B. Awerbuch, R. Khandekar, and S. Rao. Distributed algorithms for multicommodity flow problems via approximate steepest descent framework. *ACM Transactions on Algorithms*, 9(1):3, 2012.

- [9] B. Awerbuch and T. Leighton. Improved approximation algorithms for the multi-commodity flow problem and local competitive routing for dynamic networks. In *Proc. 26th Ann. ACM Symp. on Theory of Computing*, pages 487–496, 1994.
- [10] S. Baswana and S. Sen. A simple and linear time randomized algorithm for computing sparse spanners in weighted graphs. *Random Struc. & Algorithms*, 30(4):532–563, 2007.
- [11] A. A. Benczúr and D. R. Karger. Randomized approximation schemes for cuts and flows in capacitated graphs. *SIAM J. on Computing*, 44(2):290–319, 2015.
- [12] G. E. Blelloch, A. Gupta, I. Koutis, G. L. Miller, R. Peng, and K. Tangwongsan. Nearly-linear work parallel SDD solvers, low-diameter decomposition, and low-stretch subgraphs. *Theory Comput. Syst.*, 55(3):521–554, 2014.
- [13] P. Christiano, J. A. Kelner, A. Madry, D. A. Spielman, and S.-H. Teng. Electrical flows, laplacian systems, and faster approximation of maximum flow in undirected graphs. In *Proc. of the Symp. on Theory of Comp. (STOC)*, pages 273–282, 2011.
- [14] A. Das Sarma, S. Holzer, L. Kor, A. Korman, D. Nanongkai, G. Pandurangan, D. Peleg, and R. Wattenhofer. Distributed verification and hardness of distributed approximation. In *Proc. of the Symp. on Theory of Comp. (STOC)*, pages 363–372, 2011.
- [15] A. V. Goldberg and R. E. Tarjan. Efficient maximum flow algorithms. *Commun. ACM*, 57(8):82–89, August 2014.
- [16] J. A. Kelner, Y. T. Lee, L. Orecchia, and A. Sidford. An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations. In *SODA*, pages 217–226, 2014.
- [17] I. Koutis. Simple parallel and distributed algorithms for spectral graph sparsification. In *the Proceedings of the Symposium on Parallel Algorithms and Architectures*, pages 61–66, 2014.
- [18] S. Kutten and D. Peleg. Fast distributed construction of k-dominating sets and applications. In *the Proc. of the Int’l Symp. on Princ. of Dist. Comp. (PODC)*, pages 238–251, 1995.
- [19] A. Madry. Fast approximation algorithms for cut-based problems in undirected graphs. In *Proc. of the Symp. on Found. of Comp. Sci. (FOCS)*, pages 245–254, 2010.
- [20] A. Madry. Fast approximation algorithms for cut-based problems in undirected graphs. *CoRR*, abs/1008.1975, 2010.
- [21] J. M. Marberg and E. Gafni. An $O(n^2m^{1/2})$ distributed max-flow algorithm. In *Int. Conf. on Parallel Processing, (ICPP’87)*, pages 213–216, 1987.
- [22] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004.
- [23] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.
- [24] D. Peleg. *Distributed Computing: A Locality-Sensitive Approach*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

- [25] S. A. Plotkin, D. B. Shmoys, and É. Tardos. Fast approximation algorithms for fractional packing and covering problems. *Mathematics of Operations Research*, 20(2):257–301, 1995.
- [26] H. Räcke. Optimal hierarchical decompositions for congestion minimization in networks. In *Proc. of the Symp. on Theory of Comp. (STOC)*, pages 255–264, 2008.
- [27] A. Schrijver. On the history of the transportation and maximum flow problems. *Mathematical Programming*, 91(3):437–445, 2002.
- [28] A. Segall. Decentralized maximum-flow protocols. *Networks*, 12(3):213–230, Fall 1982.
- [29] J. Sherman. Nearly maximum flows in nearly linear time. In *Proc. of the Symp. on Found. of Comp. Sci. (FOCS)*, pages 263–269, 2013.
- [30] D. A. Spielman and S. Teng. Nearly-linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *CoRR*, abs/cs/0607105, 2006.
- [31] N. E. Young. Sequential and parallel algorithms for mixed packing and covering. In *Proc. of the Symp. on Found. of Comp. Sci. (FOCS)*, pages 538–546. IEEE, 2001.