

# Geometric Partial Matchings and Unbalanced Transportation Problem

Pankaj K. Agarwal

Duke University, USA

[pankaj@cs.duke.edu](mailto:pankaj@cs.duke.edu)

Hsien-Chih Chang

Duke University, USA

[hsienchih.chang@duke.edu](mailto:hsienchih.chang@duke.edu)

Allen Xiao

Duke University, USA

[axiao@cs.duke.edu](mailto:axiao@cs.duke.edu)

## 1 Abstract

Let  $A$  and  $B$  be two point sets in the plane with uneven sizes  $r$  and  $n$  respectively (assuming  $r$  is at most  $n$ ), and let  $k$  be a parameter. The geometric partial matching problem asks to find the minimum-cost size- $k$  matching between  $A$  and  $B$  under powers of  $L_p$  distances. Applying combinatorial algorithms for partial matching in general graphs to our setting naïvely requires quadratic time due to existence of many edges between point sets  $A$  and  $B$ . Most previous work for geometric matching has focused on the setting when  $k$ ,  $r$ , and  $n$  are all equal. The best algorithm in this setting, due to Sharathkumar and Agarwal [STOC 2012], runs in time  $O(n \text{ polylog } n \cdot \text{poly } \varepsilon^{-1})$ , but is limited to matching objectives that are sum-of-distances.

We present the first set of geometric algorithms which work for any powers of  $L_p$ -norm matching objective: An exact algorithm which runs in  $O((n + k^2) \text{ polylog } n)$  time, and a  $(1 + \varepsilon)$ -approximation which runs in  $O((n + k\sqrt{k}) \text{ polylog } n \cdot \log \varepsilon^{-1})$  time. Both algorithms are based on primal-dual flow augmentation scheme; the main improvements are obtained by using dynamic data structures to achieve efficient flow augmentations. Using similar techniques, we give an exact algorithm for the planar transportation problem, which runs in  $O(rn^{3/2} \text{ polylog } n)$  time. This is the first sub-quadratic time exact algorithm when  $r = o(\sqrt{n})$ , which improves over the state-of-art quadratic time algorithm by Agarwal *et al.* [SOCG 2016].

**2012 ACM Subject Classification** Theory of computation → Design and analysis of algorithms

**Keywords and phrases** partial matching, transportation, minimum-cost flow, rms-distance, bichromatic closest pair, cost scaling, excess scaling, primal-dual

**Lines** 1141

## 18 1 Introduction

### 19 «REWRITE AFTER TECHNICAL SECTIONS»

Consider the problem of finding a minimum-cost bichromatic matching between a set of red points  $A$  and a set of blue points  $B$  lying in the plane, where the cost of a matching edge  $(a, b)$  is the Euclidean distance  $\|a - b\|$ ; in other words, the minimum-cost bipartite matching problem on the Euclidean complete graph  $G = (A \cup B, A \times B)$ . Let  $r := |A|$  and  $n := |B|$ . Without loss of generality, assume that  $r \leq n$ . We consider the problem of *partial matching* (also called *imperfect matching*), where the task is to find a minimum-cost matching of size  $k \leq r$ . When  $k = r = n$ , we say the matching instance is *balanced*. When  $k = r < n$  ( $A$  and  $B$  have different sizes, but the matching is maximal), we say the matching instance is



© Pankaj K. Agarwal, Hsien-Chih Chang, Allen Xiao;  
licensed under Creative Commons License CC-BY

The 35th International Symposium on Computational Geometry (SOCG 2019).



Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



28 *unbalanced*. We call the geometric problem of finding a size  $k$  matching of point sets  $A$  and  
 29  $B$  the *geometric partial matching problem*. **«talk about the near-linear time algorithm»**

### 30 1.1 Contributions

31 In this paper, we present two algorithms for geometric partial matching that are based on  
 32 fitting nearest-neighbor (NN) and geometric closest pair (BCP) oracles into primal-dual  
 33 algorithms for non-geometric bipartite matching and minimum-cost flow. This pattern is  
 34 not new, see for example (...) **«TODO cite»**. Unlike these previous works, we focus on  
 35 obtaining running time dependencies on  $k$  or  $r$  instead of  $n$ , that is, faster for inputs with  
 36 small  $r$  or  $k$ . We begin in Section ?? by introducing notation for matching and minimum-cost  
 37 flow.

38 First in Section 2, we show that the Hungarian algorithm [7] combined with a BCP  
 39 oracle solves geometric partial matching exactly in time  $O((n + k^2) \text{polylog } n)$ . Mainly, we  
 40 show that we can separate the  $O(n \text{polylog } n)$  preprocessing time for building the BCP data  
 41 structure from the augmenting paths' search time, and update duals in a lazy fashion such  
 42 that the number of dual updates per augmenting path is  $O(k)$ .

43 **► Theorem 1.1.** *Let  $A$  and  $B$  be two point sets in the plane with  $|A| = r$  and  $|B| = n$   
 44 satisfying  $r \leq n$ , and let  $k$  be a parameter. A minimum-cost geometric partial matching of  
 45 size  $k$  can be computed between  $A$  and  $B$  in  $O((n + k^2) \text{polylog } n)$  time.*

46 **«State the settings separately so no need to repeat in the theorem statement.»**

47 Next in Section 3, we apply a similar technique to the unit-capacity min-cost circulation  
 48 algorithm of Goldberg, Hed, Kaplan, and Tarjan [4]. The resulting algorithm finds a  $(1 + \varepsilon)$ -  
 49 approximation to the optimal geometric partial matching in  $O((n + k\sqrt{k}) \text{polylog } n \log(n/\varepsilon))$   
 50 time.

51 **► Theorem 1.2.** *Let  $A$  and  $B$  be two point sets in the plane with  $|A| = r$  and  $|B| = n$   
 52 satisfying  $r \leq n$ , and let  $k$  be a parameter. A  $(1 + \varepsilon)$  geometric partial matching of size  $k$   
 53 can be computed between  $A$  and  $B$  in  $O((n + k\sqrt{k}) \text{polylog } n \log(1/\varepsilon))$  time.*

54 Our third algorithm solves the transportation problem in the unbalanced setting. The  
 55 transportation problem is a weighted generalization of the matching problem. Let  $A$  and  $B$   
 56 be point sets in the plane. Let  $\lambda : A \cup B \rightarrow \mathbb{Z}$  be a *supply-demand function* with positive value  
 57 on points of  $A$ , negative value on points of  $B$ , satisfying  $\sum_{a \in A} \lambda(a) = -\sum_{b \in B} \lambda(b)$ . Define  
 58  $U := \max_{p \in A \cup B} |\lambda(p)|$ . A *transportation map* is a function  $\tau : A \times B \rightarrow \mathbb{R}_{\geq 0}$ . A transportation  
 59 map  $\tau$  is *feasible* if  $\sum_{b \in B} \tau(a, b) = \lambda(a)$  for all  $a \in A$ , and  $\sum_{a \in A} \tau(a, b) = -\lambda(b)$  for all  
 60  $b \in B$ . In other words, the value  $\tau(a, b)$  describes how much supply at  $a$  should be sent to  
 61 meet demands at  $b$ , and we require that all supplies are sent and all demands are met. We  
 62 define the *cost* of  $\tau$  to be

$$63 \quad \text{cost}(\tau) := \sum_{(a,b) \in A \times B} \|a - b\|_p^q \cdot \tau(a, b).$$

64 Given  $A$ ,  $B$ , and  $\lambda$ , the *transportation problem* asks to find a feasible transportation map of  
 65 minimum cost. We focus on the case when the point sets are *unbalanced*; that is, the given  
 66 point sets have different sizes  $r$  and  $n$ ; without loss of generality assuming  $r \leq n$ . For this,  
 67 we use the strongly polynomial uncapacitated min-cost flow algorithm by Orlin [8]. The  
 68 result is an  $O(n^{3/2}r \text{polylog } n)$  time algorithm for unbalanced transportation. This improves  
 69 over the  $O(n^2 \text{polylog } n)$  time algorithm of Agarwal *et al.* [1] when  $r = o(\sqrt{n})$ .

70 ► **Theorem 1.3.** *Let  $A$  and  $B$  be two point sets in the plane with  $|A| = r$  and  $|B| = n$*   
 71 *satisfying  $r \leq n$ , with supplies and demands given by the function  $\lambda : (A \cup B) \rightarrow \mathbb{Z}$*   
 72 *such that  $\sum_{a \in A} \lambda(a) = \sum_{b \in B} \lambda(b)$ . An optimal transportation map can be computed in*  
 73  *$O(rn^{3/2} \text{polylog } n)$  time.*

74 By nature of the BCP/NN oracles we use, these results generalize to when  $\|a - b\|$  is any  
 75  $L_p$  distance, and if we use  $p'$ -th power costs  $c(a, b) = \|a - b\|^{p'}$  for any  $1 \leq p' < \infty$ .

## 76 2 Minimum-Cost Partial Matchings using Hungarian Algorithm

77 The Hungarian algorithm [7] is a primal-dual algorithm for min-cost bipartite matching in  
 78 general graphs that can be adapted to solve the partial matching problem exactly if one  
 79 terminates the algorithm after  $k$  iterations (see e.g. [9]). In this section, we prove Theorem 1.1  
 80 by implementing the Hungarian algorithm in  $O((n + k^2) \text{polylog } n)$  time.

### 81 2.1 Matching Terminologies

82 «**Move some to intro**» Let  $G$  be a bipartite graph between vertex sets  $A$  and  $B$  and edge  
 83 set  $E$ , with costs  $c(v, w)$  for each edge  $(v, w)$  in  $G$ . A *matching*  $M \subseteq E$  is a set of edges  
 84 where no two edges share an endpoint. A vertex  $v$  is *matched* by  $M$  if  $v$  is the endpoint  
 85 of some matching edge in  $M$ ; otherwise  $v$  is *unmatched*. The *size* of a matching is the  
 86 number of edges in the set, and the *cost* of a matching is the sum of costs of its edges. For  
 87 a parameter  $k$ , the *minimum-cost partial matching problem (MPM)* asks to find a size- $k$   
 88 matching of minimum cost. In the geometric partial matching setting, we have  $E = A \times B$   
 89 and  $c(a, b) = \|a - b\|_p^q$  for every edge  $(a, b)$  in  $G$ .

90 The linear program dual to the standard linear program for MPM has dual variables for  
 91 each vertex, called *potentials*  $\pi$ . Given potentials  $\pi$ , we can define the *reduced cost* on the  
 92 edges to be  $c_\pi(v, w) := c(v, w) - \pi(v) + \pi(w)$ . Potentials  $\pi$  are *feasible* if the reduced costs  
 93 are nonnegative for all edges in  $G$ . We say that an edge  $(v, w)$  is *admissible* under potentials  
 94  $\pi$  if  $c_\pi(v, w) = 0$ .

95 Consider a matching  $M$  of size less than  $r$ . An *augmenting path*  $\Pi = (a_1, b_1, \dots, a_\ell, b_\ell)$   
 96 is an odd-length path with unmatched endpoints  $(a_1$  and  $b_\ell)$  and all other points matched.  
 97 The edges of  $\Pi$  alternate between edges outside and inside of matching  $M$ . The symmetric  
 98 difference  $M \oplus \Pi$  creates a new matching of size  $|M| + 1$ . We say that  $M \oplus \Pi$  is the result of  
 99 *augmenting*  $M$  by  $\Pi$ .

### 100 2.2 The Hungarian Algorithm

101 The Hungarian algorithm is initialized with  $M = \emptyset$  and  $\pi = 0$ . Each iteration of the  
 102 Hungarian algorithm augments  $M$  with an admissible augmenting path  $\Pi$ , discovered using  
 103 a procedure called the *Hungarian search*. The algorithm terminates once  $M$  has size  $k$ ;  
 104 Ramshaw and Tarjan [9] showed that  $M$  is guaranteed to be an optimal partial matching.

105 «**For the specific implementation of Hungarian search where in each iteration the roots**  
 106 **are all unmatched vertices in  $A$ . Picking one as source does not guarantee min-cost**  
 107 **partial matching.**»

108 The Hungarian search tries grow a set of *reachable vertices*  $S$  by augmenting paths  
 109 consisting of admissible edges. Initially,  $S$  is the set of unmatched vertices in  $A$ . Let the  
 110 *frontier* of  $S$  be the edges in  $(A \cap S) \times (B \setminus S)$ . In each iteration, the Hungarian search  
 111 first *relaxes* the minimum-reduced-cost edge  $(a, b)$  in the frontier, raising  $\pi(a)$  by  $c_\pi(a, b)$   
 112 for all  $a \in S$  to make  $(a, b)$  admissible, and adding  $b$  into  $S$ . It is easy to verify that this

potential change preserves feasibility. If  $b$  is already matched, then we also relax the matching edge  $(a', b)$  and add  $a'$  into  $S$ . The search finishes when  $b$  is unmatched, and an admissible augmenting path now can be recovered.

The remainder of this section describes an implementation of Hungarian search that runs in  $O(k \text{ polylog } n)$  time after an  $O(n \text{ polylog } n)$  time preprocessing (see Lemma 2.1). Our implementation of the Hungarian algorithm therefore runs in  $O((n + k^2) \text{ polylog } n)$  time, which proves Theorem 1.1.

## 2.3 Fast implementation of Hungarian search

The most expensive step in augmentation is to find the minimum-reduced-cost frontier edge that needs to be relaxed — the search must “look at every edge”. In the geometric setting, we find the min-cost edge using a dynamic *bichromatic closest pair* (BCP) data structure, as observed in [1]. Given two point sets  $P$  and  $Q$  in the plane, the bichromatic closest pair are two points  $p \in P$  and  $q \in Q$  minimizing the additively weighted distance  $\|p - q\| - \omega(p) + \omega(q)$  for some real-valued vertex weights  $\omega$ . Thus, the minimum reduced-cost among the frontier edges is precisely the cost of the BCP of point sets  $P = A \cap S$  and  $Q = B \setminus S$ , with  $\omega(p) = \pi(p)$ .

«Under the assumption ... on the metric,» The state of the art dynamic BCP data structure from Kaplan *et al.* [6] supports point insertions and deletions in  $O(\text{polylog } n)$  time, and answers queries in  $O(\log^2 n)$  time. During each relaxation, we perform at most one query and add a vertex to  $S$  incurring one BCP insertion or deletion. Thus the running time for the search is  $O(k \text{ polylog } n)$ .

**Initial BCP sets by rewinding.** Recall that in the beginning of each iteration,  $S$  is initialized to the set of unmatched vertices in  $A$ , and therefore  $Q = B \setminus S$  has size  $n$  «what?». We cannot afford to take  $O(n \text{ polylog } n)$  time initialize the BCP data structure at the beginning of every Hungarian search beyond the first. However, the set of unmatched  $A$  vertices has changed by exactly one vertex since the last Hungarian search — the augmentation newly matched one vertex  $a^* \in A$ . Thus, given the initial BCP sets  $P', Q'$  from the beginning of the last Hungarian search, we can construct  $P$  and  $Q$  for the current iteration using a single BCP deletion in  $O(\text{polylog } n)$  time.

To acquire  $P'$  and  $Q'$ , we keep track of a list of the points added to  $S$  over the course of the Hungarian search. At the end of each Hungarian search we *rewind* the BCP data structure by tracing the list in reverse order. The number of points in the list is at most  $O(k)$  as it is bounded by the number of relaxations per Hungarian search. Thus, in  $O(k \text{ polylog } n)$  time, we can reconstruct  $P'$  and  $Q'$  for each Hungarian search beyond the first. We refer to this procedure as the *rewinding mechanism*.

### Potential updates. «REWRITE TO BE MORE SUCCINCT.»

We modify a trick from Vaidya [11] for batching potential updates. Potentials have a *stored value*, i.e. the currently recorded value of  $\pi(v)$ , and a *true value*, which may have changed from  $\pi(v)$ . The resulting algorithm queries the minimum-reduced-cost under the true values of  $\pi$  and updates the stored value occasionally.

Throughout the entire Hungarian algorithm, we maintain a nonnegative scalar  $\delta$  (initially set to 0) which aggregates potential changes. Vertices  $a \in A$  that are added to  $S$  are inserted into BCP with weight  $\omega(a) \leftarrow \pi(a) - \delta$ , for whatever value  $\delta$  is at the time of insertion. Similarly, vertices  $b \in B$  that are added to  $S$  have  $\omega(b) \leftarrow \pi(b) - \delta$  recorded ( $B \cap S$  points aren't added into a BCP set). When the Hungarian search wants to raise the potentials of points in  $S$ ,  $\delta$  is increased by that amount instead. Thus, true value for any potential of a

point in  $S$  is always  $\omega(p) + \delta$ . For points of  $(A \cup B) \setminus S$ , the true potential is equal to the stored potential. Since all the points of  $A \cap S$  have weights uniformly offset from their true potentials, the minimum edge returned by the BCP does not change. **«why?»**

Once a point is removed from  $S$  (i.e. by an augmentation or the rewinding mechanism), we update its stored potential  $\pi(p) \leftarrow \omega(p) + \delta$ , again for the current value of  $\delta$ . Most importantly,  $\delta$  is not reset at the end of a Hungarian search and persists through the entire algorithm. Thus, the initial BCP sets constructed by the rewinding mechanism have true potentials accurately represented by  $\delta$  and  $\omega(p)$ .

We update  $\delta$  once per edge relaxations; thus  $O(k)$  times in total per Hungarian search. There are  $O(k)$  stored values updated per Hungarian search during the rewinding process. The time spent on potential updates per Hungarian search is therefore  $O(k)$ .

Putting everything together we obtain the following:

► **Lemma 2.1.** *Each Hungarian search can be implemented in  $O(k \text{ polylog } n)$  time after a one-time  $O(n \text{ polylog } n)$  preprocessing.*

### 3 Approximating Min-Cost Partial Matching through Cost-Scaling

The goal of section is to prove Theorem 1.2; that is, to compute a size- $k$  geometric partial matching between two point sets  $A$  and  $B$  in the plane, with cost at most  $(1 + \varepsilon)$  times the optimal matching, in time  $O((n + k\sqrt{k}) \text{ polylog } n \log(1/\varepsilon))$ .

**«Summarize our new ideas that lead to the improvement.»**

After introducing the necessary terminologies in Section 3.1, we reduce the partial matching problem to computing an approximate minimum-cost flow on a unit-capacity reduction network in Section 3.2. In Section 3.3 we outline the high-level overview of the cost-scaling algorithm. We postpone the fast implementation using dynamic data structures to Section 4.

#### 3.1 Preliminaries on Network Flows

Due to the space restriction, we omit the definitions of standard network flow theory terminologies from the main text. For a reference see Appendix A.1, or any texts on network flows [1]. We emphasize that a directed graph  $G = (V, E)$  is augmented by edge costs  $c$  and capacities  $u$ , and a supply-demand function  $\phi$  defined on the vertices. A **network**  $N = (V, \vec{E})$  turns each edge in  $E$  into a pair of **arcs**  $v \rightarrow w$  and  $w \rightarrow v$  in arc set  $\vec{E}$ . With the unit-capacity assumption on the network, all the pseudoflows in this section take integer values. The **support** of a pseudoflow  $f$  in  $N$ , denoted as  $\text{supp}(f)$ , is the set of arcs with positive flows:  $\text{supp}(f) := \{v \rightarrow w \in \vec{E} \mid f(v \rightarrow w) > 0\}$ . If all vertices are **balanced**, the pseudoflow is a **circulation**. The **cost** of a pseudoflow is defined to be

$$\text{cost}(f) := \sum_{v \rightarrow w \in \text{supp}(f)} c(v \rightarrow w) \cdot f(v \rightarrow w).$$

The **minimum-cost flow problem (MCF)** asks to find a circulation of minimum cost inside a given directed graph.

**LP-duality and admissibility.** To solve the minimum-cost flow problem, we focus on the primal-dual algorithms using linear programming. Let  $G = (V, E)$  be a given directed graph with the corresponding network  $N = (V, \vec{E}, c, u, \phi)$ . Formally, the **potentials**  $\pi(v)$  are the variables of the linear program dual to the standard linear program for the minimum-cost

flow problem with variables  $f(v, w)$  for each directed edge in  $E$ . Assignments to the primal variables satisfying the capacity constraints extend naturally into a pseudoflow on the network  $N$ . Let  $G_f = (V, \vec{E}_f)$  be the residual graph under pseudoflow  $f$ . The *reduced cost* of an arc  $v \rightarrow w$  in  $\vec{E}_f$  with respect to  $\pi$  is defined as  $c_\pi(v \rightarrow w) := c(v \rightarrow w) - \pi(v) + \pi(w)$ . Notice that the cost function  $c_\pi$  is also antisymmetric.

The *dual feasibility constraint* says that  $c_\pi(v \rightarrow w) \geq 0$  holds for every directed edge  $(v, w)$  in  $E$ ; potentials  $\pi$  which satisfy this constraint are said to be *feasible*. Suppose we relax the dual feasibility constraint to allow some small violation in the value of  $c_\pi(v \rightarrow w)$ . We say that a pair of pseudoflow  $f$  and potential  $\pi$  is  $\varepsilon$ -*optimal* [?, ?] if  $c_\pi(v \rightarrow w) \geq -\varepsilon$  for every residual arc  $v \rightarrow w$  in  $\vec{E}_f$ . Pseudoflow  $f$  is  $\varepsilon$ -*optimal* if it is  $\varepsilon$ -optimal with respect to some potentials  $\pi$ ; potential  $\pi$  is  $\varepsilon$ -*optimal* if it is  $\varepsilon$ -optimal with respect to some pseudoflow  $f$ . Given a pseudoflow  $f$  and potentials  $\pi$ , a residual arc  $v \rightarrow w$  in  $\vec{E}_f$  is *admissible* if  $c_\pi(v \rightarrow w) \leq 0$ . We say that a pseudoflow  $g$  in  $G_f$  is *admissible* if all support arcs of  $g$  on  $G_f$  are admissible; in other words,  $g(v \rightarrow w) > 0$  holds only on admissible arcs  $v \rightarrow w$ .

► **Lemma 3.1.** *Let  $f$  be an  $\varepsilon$ -optimal pseudoflow in  $G$  and let  $f'$  be an admissible flow in  $G_f$ . Then  $f + f'$  is also  $\varepsilon$ -optimal. <<Lemma 5.3 in [5]? Also, where is this used?>>*

### 3.2 Reduction to Unit-Capacity Min-Cost Flow Problem

The goal of the subsection is to reduce the minimum-cost partial matching problem to the unit-capacity minimum-cost flow problem with a polynomial bound on diameter of the underlying point set. To this end we first provide an upper bound on the size of support of an integral pseudoflow on the standard reduction network between the two problems. This upper bound in turn provides an additive approximation on the cost of an  $\varepsilon$ -optimal circulation. Next we employ a technique by Sharathkumar and Agarwal [10] to transform an additive  $\varepsilon$ -approximate solution into a multiplicative  $(1 + \varepsilon)$ -approximation for the geometric partial matching problem. The reduction does not work out of the box, as Sharathkumar and Agarwal were tackling a similar but different problem on geometric transportations.

► **Lemma 3.2.** *Computing a  $(1 + \varepsilon)$ -approximate geometric partial matching can be reduced to the following problem in  $O(n \text{ polylog } n)$  time: Given a reduction network  $N$  over a point set with diameter at most  $K \cdot kn^3$  for some constant  $K$ , compute a  $(K \cdot \varepsilon / 6k)$ -optimal circulation on  $N$ .*

**Additive approximation.** Given a bipartite graph  $G = (A, B, E_0)$  for the geometric partial matching problem with cost function  $c$ , we construct the *reduction network*  $N_H$  as follows: Direct the edges in  $E_0$  from  $A$  to  $B$ , and assign each directed edge with capacity 1. Now add a dummy vertex  $s$  with directed edges to all vertices in  $A$ , and add a dummy vertex  $t$  with directed edges from all vertices in  $B$ ; each edge added this way has cost 0 and capacity 1. Denote the new graph with vertex set  $V = A \cup B \cup \{s, t\}$  and edge set  $E$  as the *reduction graph*  $H$ . Assign vertex  $s$  with supply  $k$  and vertex  $t$  with demand  $k$ ; the rest of the vertices in  $H$  have zero supply-demand. We call the network naturally corresponds to  $H$  as the *reduction network*, denoted by  $N_H$ .

It is straightforward to show that any integer circulation  $f$  on  $N_H$  uses exactly  $k$  of the  $A$ -to- $B$  arcs, which correspond to the edges of a size- $k$  matching  $M_f$ . Notice that the cost of the circulation  $f$  is equal to the cost of the corresponding matching  $M_f$ .

First we show that the number of arcs used by any integer pseudoflow in  $N_H$  is asymptotically bounded by the excess of the pseudoflow.

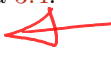


243 ► **Lemma 3.3.** *The size of  $\text{supp}(f)$  is at most  $3k$  for any integer circulation  $f$  in reduction*  
 244 *network  $N_H$ . As a corollary, the number of residual backward arcs is at most  $3k$ .*

245 Using the bound on the support size, we show that an  $\varepsilon$ -optimal integral circulation gives  
 246 an additive  $O(k\varepsilon)$ -approximation to the MCF problem.

247 ► **Lemma 3.4.** *Let  $f$  be an  $\varepsilon$ -optimal integer circulation in  $N_H$ , and  $f^*$  be an optimal integer*  
 248 *circulation for  $N_H$ . Then,  $\text{cost}(f) \leq \text{cost}(f^*) + 6k\varepsilon$ .*

249 **Multiplicative approximation.** Now we employ a technique from Sharathkumar and Agarwal [10] to convert the additive approximation into a multiplicative one. Here we sketch a proof to Lemma 3.2; a complete proof can be found in the Appendix. **«Is it okay if we hide the parametric search?»**

250 Sharathkumar and Agarwal [10, §3.5] provide a construction that partitions the input  
 251 point sets  $A$  and  $B$  for the MPM problem into clusters, such that the diameter of each cluster  
 252 is upper bounded by the cost of the optimal partial matching multiply by a polynomial  
 253 factor. To prove Lemma 3.2, we further modify the point set by moving the clusters so  
 254 that the cost of the optimal solution does not change, while the diameter of the *whole* point  
 255 set is bounded. Now one can prove Lemma 3.2 by computing an  $(\varepsilon \text{cost}(M^*)/6k)$ -optimal  
 256 circulation  $f$  on the modified point set using additive approximation from Lemma 3.4.  
 257 

### 260 3.3 High-Level Description of Cost-Scaling Algorithm

261 Our main algorithm for the unit-capacity minimum-cost flow problem is based on the *cost-*  
 262 *scaling* technique, originally due to Goldberg and Tarjan [5]; Goldberg *et al.* [4] applied  
 263 the technique on unit-capacity networks. The algorithm finds  $\varepsilon$ -optimal circulations for  
 264 geometrically shrinking values of  $\varepsilon$ . Each fixed value of  $\varepsilon$  is called a *cost scale*. Once  $\varepsilon$  is  
 265 sufficiently small, the  $\varepsilon$ -optimal flow is a suitable approximation according to Lemma 3.2.<sup>1</sup>

266 The cost-scaling algorithm initializes the flow  $f$  and the potential  $\pi$  to be zero. Note  
 267 that the zero flow is trivially a  $kC$ -optimal flow, where  $C$  is the maximum arc cost. At the  
 268 beginning of each scale starting at  $\varepsilon = kC$ ,

- 269 ■ SCALE-INIT takes the previous circulation (now  $2\varepsilon$ -optimal) and transforms it into an  
 270  $\varepsilon$ -optimal pseudoflow with  $O(k)$  excess.
- 271 ■ REFINE then reduces the excess in the newly constructed pseudoflow to zero, making it  
 272 an  $\varepsilon$ -optimal circulation.

273 Thus for any  $\varepsilon^* > 0$ , the algorithm produces an  $\varepsilon^*$ -optimal circulation after  $O(\log(kC/\varepsilon^*))$   
 274 scales. Using the reduction in Lemma 3.2, we have the diameter of the point set, thus  
 275 maximum cost  $C$ , bounded by  $O(K \cdot kn^3)$  for some value  $K$ . By setting  $\varepsilon^*$  to be  $K \cdot \varepsilon/6k$ ,  
 276 the number of cost scales is bounded above by  $O(\log(n/\varepsilon))$ .

280 **Scale initialization.** Recall that  $H$  is the *reduction graph* and  $N_H$  is the *reduction network*,  
 281 both constructed in Section 3.2. The vertex set of  $H$  consists of two point sets  $A$  and  $B$ , as  
 282 well as two dummy vertices  $s$  and  $t$ . The directed edges in  $H$  are pointed from  $s$  to  $A$ , from  
 283  $A$  to  $B$ , and from  $B$  to  $t$ . We call those arcs in  $N_H$  whose direction is consistent with their  
 284 corresponding directed edges as *forward arcs*, and those arcs that points in the opposite  
 285 direction as *backward arcs*.

261 <sup>1</sup> When the costs are integers, an  $\varepsilon$ -optimal circulation for a sufficiently small  $\varepsilon$  (say less than  $1/n$ ) is itself  
 262 an optimal solution [4, 5]. We present this algorithm without the integral-cost assumption because in  
 263 the geometric partial matching setting (with respect to  $L_p$  norms) the costs are generally not integers.

*different  
from  
original*

The procedure SCALE-INIT transforms a  $2\varepsilon$ -optimal circulation from the previous cost scale into an  $\varepsilon$ -optimal flow with  $O(k)$  excess, by raising the potentials  $\pi$  of all vertices in  $A$  by  $\varepsilon$ , those in  $B$  by  $2\varepsilon$ , and the potential of  $t$  by  $3\varepsilon$ . The potential of  $s$  remains unchanged. Now the reduced cost of every forward arc is dropped by  $\varepsilon$ , and thus all the forward arcs have reduced cost at least  $-\varepsilon$ .

As for backward arcs, the procedure SCALE-INIT continues by setting the flow on  $v \rightarrow w$  to zero for each backward arc  $w \rightarrow v$  violating the  $\varepsilon$ -optimality constraint. In other words, we set  $f(v \rightarrow w) = 0$  whenever  $c_\pi(w \rightarrow v) < -\varepsilon$ . This ensures that all such backward arcs are no longer residual, and therefore the flow (now with excess) is  $\varepsilon$ -optimal.

Because the arcs are of unit-capacity in  $N_H$ , each arc desaturation creates one unit of excess. By Lemma 3.3 the number of backward arcs is at most  $3k$ . Thus the total amount of excess created is also  $O(k)$ .

In total, potential updates and backward arc desaturations, thus the whole procedure SCALE-INIT, take  $O(n)$  time.

**Refinement.** The procedure REFINE is implemented using a primal-dual augmentation algorithm, which sends flows on admissible arcs to reduce the total excess, like the Hungarian algorithm. Unlike the Hungarian algorithm, it uses *blocking flows* instead of augmenting paths. We call a pseudoflow  $f$  on residual network  $N_g$  a *blocking flow* if  $f$  saturates at least one residual arc in every augmenting path in  $N_g$ . In other words, there is no admissible augmenting path in  $N_{f+g}$  from an excess vertex to a deficit vertex.

Each iteration of REFINE finds an admissible blocking flow that is then added to the current pseudoflow in two stages:

1. A *Hungarian search*, which increases the dual variables  $\pi$  of vertices that are reachable from an excess vertex by at least  $\varepsilon$ , in a Dijkstra-like manner, until there is an excess-deficit path of admissible edges.
2. A *depth-first search* through the set of admissible edges to construct an admissible blocking flow. It suffices to repeatedly extract admissible augmenting paths until no more admissible excess-deficit paths remain. By definition, the union of such paths is a blocking flow.

The algorithm continues until the total excess becomes zero and the  $\varepsilon$ -optimal flow is now a circulation.

First we analyze the number of iterations executed by REFINE. The proof follows the strategy in Goldberg *et al.* [4, Section 3.2]. Due to space constraint we omit all the proofs here; see Appendix A.3 for complete proofs. **«Explain what is new here.»**

► **Lemma 3.5.** *Let  $f$  be a pseudoflow in  $N_H$  with  $O(k)$  excess. The procedure REFINE runs for  $O(\sqrt{k})$  iterations before the excess of  $f$  becomes zero.*

The goal of the next section is to show that after  $O(n \text{ polylog } n)$  time preprocessing, each Hungarian search and depth-first search can be implemented in  $O(k \text{ polylog } n)$  time. Combined with the  $O(\sqrt{k})$  bound on the number of iterations we just proved, the procedure REFINE can be implemented in  $O((n + k\sqrt{k}) \text{ polylog } n)$  time. Together with our analysis on scale initialization and the bound on number of cost scales, this concludes the proof to Theorem 1.2.

## 4 Fast Implementation of Refinement

Both Hungarian search and depth-first search are implemented in a Dijkstra-like fashion, traversing through the residual graph using admissible arcs starting from the excess vertices.

Describing  
to the  
same.

△



Each step of the search procedures *relaxes* a minimum-reduced-cost arc from the set of visited vertices to an unvisited vertex, until a deficit vertex is reached. At a high level, our analysis strategy is to charge the relaxation events to the support arcs of  $f$ , which has size at most  $O(k)$  by Corollary A.3.

## 4.1 Null vertices and shortcut graph

As it turns out, there are some vertices visited by a relaxation event which we cannot charge to  $\text{supp}(f)$ . Unfortunately the number of such vertices can be as large as  $\Omega(n)$ . To overcome this issue, we replace the residual graph with an equivalent graph that excludes all the null vertices, and run the Hungarian search and depth-first search on the resulting graph instead.

**Null vertices.** We say a vertex  $v$  in the residual graph  $N_f$  is a *null vertex* if  $\phi_f(v) = 0$  and no arcs of  $\text{supp}(f)$  is incident to  $v$ . We use  $A_\emptyset$  and  $B_\emptyset$  to denote the null vertices  $A$  and  $B$  respectively. Vertices that are not null are called *normal vertices*. A *null 2-path* is a length-2 subpath in  $N_f$  from a normal vertex to another normal vertex, passing through a null vertex. As every vertex in  $A$  has in-degree 1 and every vertex in  $B$  has out-degree 1 in the residual graph, the null 2-paths must be of the form either  $(s, v, b)$  for some vertex  $b$  in  $B \setminus B_\emptyset$  or  $(a, v, t)$  for some vertex  $a$  in  $A \setminus A_\emptyset$ . In either case, we say that the null 2-path *passes through* null vertex  $v$ . Similarly, we define the length-3 path from  $s$  to  $t$  that passes through two null vertices to be a *null 3-path*. Because reduced costs telescope for residual paths, the reduced cost of any null 2-path or null 3-path does not depend on the null vertices it passes through.

**Shortcut graph.** We construct the *shortcut graph*  $\tilde{H}_f$  from the reduction network  $H$  by removing all null vertices and their incident edges, followed by inserting an arc from the head of each null path  $\Pi$  to its tail, with cost equals to the sum of costs on the arcs. We call this arc the *shortcut* of null path  $\Pi$ , denoted as  $\text{short}(\Pi)$ . The resulting multigraph  $\tilde{H}_f$  contains only normal vertices of  $H_f$ , and the reduced cost of any path between normal vertices are preserved. We argue now that  $\tilde{H}_f$  is fine as a surrogate for  $H_f$ . Let  $\tilde{\pi}$  be an  $\varepsilon$ -optimal potential on  $\tilde{H}_f$ . Construct potentials  $\pi$  on  $H_f$  which extends  $\tilde{\pi}$  to null vertices, by setting  $\pi(a) := \tilde{\pi}(s)$  for  $a \in A_\emptyset$  and  $\pi(b) := \tilde{\pi}(t)$  for  $b \in B_\emptyset$ .

► **Lemma 4.1.** *Consider  $\tilde{\pi}$  an  $\varepsilon$ -optimal potential on  $\tilde{H}_f$  and  $\pi$  the corresponding potential constructed on  $H_f$ . Then,*

1. *potential  $\pi$  is  $\varepsilon$ -optimal on  $H_f$ , and*
2. *if arc  $\text{short}(\Pi)$  is admissible under  $\tilde{\pi}$ , then every arc in  $\Pi$  is admissible under  $\pi$ .*

«Talk about the size of shortcut graph briefly.»

## 4.2 Dynamic data structures for search procedures

**Hungarian search.** Conceptually, we are executing the Hungarian search on the shortcut graph  $\tilde{H}_f$ . We describe how we can query the minimum-reduced-cost arc leaving  $\tilde{S}$  in  $O(\text{polylog } n)$  time for the shortcut graph, without constructing  $\tilde{H}_f$  explicitly. For this purpose, let  $S$  be a set of “reached” vertices maintained, identical to  $\tilde{S}$  except whenever a shortcut is relaxed, we add the null vertices passed by the corresponding null path to  $S$  in addition to its (normal) endpoints. Observe that the arcs of  $\tilde{H}_f$  leaving  $\tilde{S}$  fall into  $O(1)$  categories:

- non-shortcut backward arcs  $(v, w)$  with  $(w, v) \in \text{supp}(f)$ ;
- non-shortcut  $A$ -to- $B$  forward arcs;

- 373 ■ non-shortcut forward arcs from  $s$ -to- $A$  and from  $B$ -to- $t$ ;
  - 374 ■ shortcut arcs  $(s, b)$  corresponding to null 2-paths from  $s$  to  $b \in (B \setminus B_\emptyset) \setminus S$ ;
  - 375 ■ shortcut arcs  $(a, t)$  corresponding to null 2-paths from  $a \in (A \setminus A_\emptyset) \cap S$  to  $t$ ; and
  - 376 ■ shortcut arcs  $(s, t)$  corresponding to null 3-paths.
- 377 For ~~arcs in~~ each category we maintain a proper data structure (either heap or BCP) to query  
 378 the min-cost arc. *repetitive arcs*

379 **Depth-first search.** Depth-first search is similar to Hungarian search in that it uses the  
 380 relaxation of minimum-reduced-cost arcs/null paths, this time to identify admissible arcs/null  
 381 paths in a depth-first manner. Similar to the Hungarian search, for each category of arcs  
 382 in  $\tilde{H}_f$  leaving  $\tilde{S}$ , we maintain a proper data structure to answer the minimum-reduced-cost  
 383 arc leaving a *fixed* vertex in  $\tilde{S}$  given by the query. Thus unlike Hungarian search which uses  
 384 BCP data structures, we use dynamic nearest-neighbor data structures instead [].

385 Each of the above data structures requires  $O(1)$  queries and updates per relaxation. So  
 386 in collaboration each relaxation can be implemented in  $O(\text{polylog } n)$  time [?, ?].

387 **Time analysis.** The complete time analysis can be found in Appendix B.2 and B.3; here we  
 388 sketch the ideas. First we show (in Appendix B.2) that both Hungarian search and depth-first  
 389 search performs  $O(k)$  relaxations before a deficit vertex is reached, by looking at shortcut  
 390 and non-shortcut relaxations separately. Both types of relaxations are eventually charged  
 391 to the support size of  $f$ . As for the time analysis, *(B.3)* using the same rewinding mechanism as  
 392 in Section 2.3, the running time of the Hungarian search and depth-first search, other than  
 393 the potential updates, can be charged to the number of relaxations. Again using the trick  
 394 by Vaidya [11] in Section 2.3 we can charge the potential updates of normal vertices to the  
 395 number of relaxations in the Hungarian search. We never explicitly maintain the potentials  
 396 on the null vertices; instead they are reconstructed whenever needed, either at the end of  
 397 each iteration of refinement or when an augmentation sends flow through a null vertex. We  
 398 show that such updates does not happen often in Appendix B.4. This completes the time  
 399 analysis, which we summarize as follows:

400 ► **Lemma 4.2.** *After  $O(n \text{ polylog } n)$ -time preprocessing, each Hungarian search and depth-*  
 401 *first search can be implemented in  $O(k \text{ polylog } n)$  time.*

## 402 5 Unbalanced Transportation

403 In this section, we give an exact algorithm which solves the planar transportation problem in  
 404  $O(rn^{3/2} \text{ polylog } n)$  time, proving Theorem 1.3. Our strategy is to use the standard reduction  
 405 to the uncapacitated min-cost flow problem, and provide a fast implementation under the  
 406 geometric setting for the uncapacitated min-cost flow algorithm by Orlin [8], combined with  
 407 some of the tools developed in Sections 2 and 3.

### 408 5.1 Uncapacitated MCF by excess scaling

409 We give an outline of the strongly polynomial algorithm for uncapacitated min-cost flow  
 410 problem from Orlin [8]. Orlin's algorithm follows an *excess-scaling* paradigm originally due  
 411 to Edmonds and Karp [2]. Consider the basic primal-dual framework used in the previous  
 412 sections: The algorithm begins with both flow  $f$  and potentials  $\pi$  set to zero. Repeatedly  
 413 runs a *Hungarian search* that raises potentials (while maintaining dual feasibility) to create  
 414 an admissible augmenting excess-deficit path, on which we perform flow augmentations. In

terms of cost,  $f$  is maintained to be 0-optimal with respect to  $\pi$  and each augmentation over admissible edges preserve such property by Lemma 3.1. Thus, the final circulation must be optimal. The excess-scaling paradigm builds on top of this skeleton by specifying (i) between which excess and deficit vertices we send flows, and (ii) how much flow is sent by the augmentation.

The excess-scaling algorithm maintains a *scale parameter*  $\Delta$ , initially set to  $U$ . A vertex  $v$  with  $|\phi_f(v)| \geq \Delta$  is called *active*. Each augmenting path is chosen between an active excess vertex and an active deficit vertex. Once there are no more active excess or deficit vertices,  $\Delta$  is halved. Each sequence of augmentations where  $\Delta$  holds a constant value is called an *excess scale*. There are  $O(\log U)$  excess scales before  $\Delta < 1$  and, by integrality of supplies/demands,  $f$  is a circulation.

With some modifications to the excess-scaling algorithm, Orlin [8] obtains a strongly polynomial bound on the number of augmentations and excess scales. First, an *active* vertex is redefined to be one satisfying  $|\phi_f(v)| \geq \alpha\Delta$ , for a fixed parameter  $\alpha \in (0.5, 1)$ . Second, arcs with flow value at least  $3n\Delta$  at the beginning of a scale are *contracted* to create a new vertex, whose supply-demand is the sum of those on the two endpoints of the contracted arc. We use  $\hat{G} = (\hat{V}, \hat{E})$  to denote the resulting *contracted graph*, where each  $\hat{v} \in \hat{V}$  is a contracted component of vertices from  $V$ . Intuitively, the flow is so high on contracted arcs that no set of future augmentations can remove the arc from  $\text{supp}(f)$ . Third, in addition to halving,  $\Delta$  is aggressively lowered to  $\max_{v \in V} \phi_f(v)$  if there are no active excess vertices and  $f(v \rightarrow w) = 0$  holds for every arc  $v \rightarrow w \in \hat{E}$ . Finally, flow values are not tracked within contracted components, but once an optimal circulation is found on  $\hat{G}$ , optimal potentials  $\pi^*$  can be *recovered* for  $G$  by sequentially undoing the contractions. The algorithm then performs a post-processing step which finds the optimal circulation  $f^*$  on  $G$  by solving a max-flow problem on the set of admissible arcs under  $\pi^*$ .

► **Theorem 5.1 (Orlin [8, Theorems 2 and 3]).** *Orlin’s algorithm finds a set of optimal potentials after  $O(n \log n)$  scaling phases and  $O(n \log n)$  total augmentations.*

The remainder of the section focuses on showing that each augmentation can be implemented in  $O(r\sqrt{n} \text{polylog } n)$  time (after preprocessing). Additionally, we show that  $f^*$  can be recovered from  $\pi^*$  very quickly in our setting.

**Implementing contractions.** **«REWRITE»** Following Agarwal *et al.* [1], our geometric data structures must deal with real points in the plane instead of the the contracted components. We will track the contracted components described in  $\hat{G}$  (e.g. with a disjoint-set data structure) and mark the arcs of  $\text{supp}(f)$  that are contracted. We maintain potentials on the points  $A$  and  $B$  directly, instead of the contracted components.

When conducting the Hungarian search, we initialize  $S$  to be the set of vertices from *active excess contracted components* who (in sum) meet the imbalance criteria. **«unclear»** Upon relaxing any  $v \in \hat{v}$ , we immediately relax all the contracted support arcs which span  $\hat{v}$ . Since the input network is uncapacitated, each contracted component is strongly connected in the residual network by the admissible forward/backward arcs of each contracted arc. **«unparsable»** To relax arcs in  $\hat{E}$ , we relax the support arcs before attempting to relax any non-support arcs. **«mention the reason to make support acyclic»** Relaxations of support arcs can be performed without further potential changes, since they are admissible by invariant.

During the augmentations, contracted residual arcs are considered to have infinite capacity, and we do not update the value of flows on these arcs. We allow augmenting paths to begin

from any point  $a \in \hat{v} \cap A$  in an active excess component  $\hat{v}$ , and end at any point  $b \in \hat{w} \cap B$  in an active deficit component  $\hat{w}$ .

**Recovering optimal flow.** Rewinding contracted components to recover an optimal flow naïvely takes  $O(rn^2)$  time. Use a strategy from Agarwal *et al.* [1], we can recover the optimal flow in time  $O(n \text{ polylog } n)$ . If furthermore the coinadmissible is just the  $p$ -norm (without the  $q$ th-power), an even stronger result stands: In this case, the set of admissible arcs under an optimal potential forms a planar graph, and thus we can apply the planar maximum-flow algorithm [?, 3] which runs in  $O(n \log n)$  time. For details see the appendix.

## 5.2 Dead vertices and support stars

Our goal is to implement each augmentation in  $O(r\sqrt{n} \text{ polylog } n)$  time. To find an augmenting path, we again use a Hungarian search with geometric data structures to perform relaxations quickly. Like in Section 3, there are vertices which cannot be charged to the flow support. Even worse, the flow support for the transportation problem may have size  $\Omega(n)$  (consider when  $A$  has one point, and demands are uniformly distributed among the vertices of  $B$ ). Our strategy is summarized as follows:

- Discard vertices which lead to dead ends in the search (not on a path to a deficit vertex).
- Cluster parts of the flow support, such that the number of support arcs outside clusters is  $O(r)$ . The number of relaxations we perform is proportional to the number of support arcs outside of clusters.

Querying/updating clusters degrades our amortized time per relaxation from  $O(\text{polylog } n)$  to  $O(\sqrt{n} \text{ polylog } n)$ . Thus overall each augmentation takes  $O(r\sqrt{n} \text{ polylog } n)$  time.

**Dead vertices.** Let the *support degree* of a vertex be its degree in the graph induced by the underlying edges of  $\text{supp}(f)$ . We call a vertex  $b \in B$  *dead* if  $b$  has support degree 0 and is not an active excess or deficit vertex; call it *live* otherwise. Dead vertices are essentially equivalent to the *null vertices* of Section 3. However, since the reduction in this section does not use a super-source/super-sink, we can simply remove these from consideration during a Hungarian search — they will not terminate the search, and have no outgoing residual arcs. Like the null vertices, we ignore dead vertex potentials and infer feasible potentials when they become live again. We use  $A_\ell$  and  $B_\ell$  to denote the living vertices of points in  $A$  and  $B$ , respectively. Note that being dead/alive is a notion strictly defined only for vertices, and not for contracted components.

We say a dead vertex is *revived* when it stops meeting either condition of the definition. Dead vertices are only revived after  $\Delta$  decreases (at the start of a subsequent excess scale) as no augmenting path will cross a dead vertex and they cannot meet the criteria for contractions. When a dead vertex is revived, we must add it back into each of our data structures and give it a feasible potential. For revived  $b \in B$ , a feasible choice of potential is  $\pi(b) \leftarrow \max_{a \in A} (\pi(a) - c(a, b))$  which we can query by maintaining a weighted nearest neighbor data structure on the points of  $A$ . The total number of revivals is bounded above by the number of augmentations: since the final flow is a circulation on  $\hat{G}$  and a newly revived vertex  $v$  has no incident arcs in  $\text{supp}(f)$  and cannot be contracted, there is at least one subsequent augmentation which uses  $v$  as its beginning or end. Thus, the total number of revivals is  $O(n \log n)$ .

*Condense*

**Support stars.** The vertices of  $B$  with support degree 1 are partitioned into subsets  $\Sigma_a \subset B$  by the  $a \in A$  lying on the other end of their single support arc. We call  $\Sigma_a$  the *support star* centered at  $a \in A$ .

Roughly speaking, we would like to handle each support star as a single unit. When the Hungarian search reaches  $a$  or any  $b \in \Sigma_a$ , the entirety of  $\Sigma_a$  (as well as  $a$ ) is also admissibly-reachable and can be included into  $S$  without further potential updates. Additionally, the only outgoing residual arcs of every  $b \in \Sigma_a$  lead to  $a$ , thus the only way to leave  $\Sigma_a \cup \{a\}$  is through an arc leaving  $a$ . Once a relaxation step reaches some  $b \in \Sigma_a$  or  $a$  itself, we would like to quickly update the state such that the rest of  $b \in \Sigma_a$  is also reached without performing relaxation steps to each individual  $b \in \Sigma_a$ .

### 5.3 Implementation details

Before describing our workaround for support stars, we analyze the number of relaxation steps for arcs outside of support stars.

► **Lemma 5.2.** *Suppose we have stripped the graph of dead vertices. The number of relaxation steps in a Hungarian search outside of support stars is  $O(r)$ .*

The running time of a Hungarian search will be  $O(r)$  times the time it takes us to implement each relaxation.

**Relaxations outside support stars.** For relaxations that don't involve support star vertices, we can once again maintain a BCP to query the minimum  $A_\ell$ -to- $B_\ell$  arc. To elaborate, this is the BCP between  $P = A_\ell \cap S$  and  $Q = (B_\ell \setminus (\bigcup_{a \in A_\ell} \Sigma_a)) \setminus S$ , weighted by potentials. This can be queried in  $O(\log^2 n)$  time and updated in  $O(\text{polylog } n)$  time per point. Since it doesn't deal with support stars, there is at most one insertion/deletion per relaxation step.

For  $B_\ell$ -to- $A_\ell$ , backward (support) arcs are kept admissible by invariant, so we relax them immediately when they arrive on the frontier.

**Relaxing a support star.** We classify support stars into two categories: *big stars* are those with  $|\Sigma_a| > \sqrt{n}$ , and *small stars* are those with  $|\Sigma_a| \leq \sqrt{n}$ . Let  $A_{\text{big}} \subseteq A$  denote the centers of big stars and  $A_{\text{small}} \subseteq A$  denote the centers of small stars. We keep the following data structures to manage support stars.

1. For each big star  $\Sigma_a$ , we use a data structure  $\mathcal{D}_{\text{big}}(a)$  to maintain BCP between  $P = A_\ell \cap S$  and  $Q = \Sigma_a$ , weighted by potentials. We query this until  $a \in S$  or any vertex of  $\Sigma_a$  is added to  $S$ .
2. All small stars are added to a single BCP data structure  $\mathcal{D}_{\text{small}}$  between  $P = A_\ell \cap S$  and  $Q = (\bigcup_{a \in A_{\text{small}}} \Sigma_a) \setminus S$ , weighted by potentials. When an  $a \in A_{\text{small}}$  or any vertex of its support star is added to  $S$ , we remove the points of  $\Sigma_a$  from  $\mathcal{D}_{\text{small}}$  using  $|\Sigma_a|$  deletion operations.

We will update these data structures as each support star center is added into  $S$ . If a relaxation step adds some  $b \in B_\ell$  and  $b$  is in a support star  $\Sigma_a$ , then we immediately relax  $b \rightarrow a$ , as all support arcs are admissible. Relaxations of non-support star  $b \in B_\ell$  will not affect the support star data structures.

Suppose a relaxation step adds some  $a \in A_\ell$  to  $S$ . For the support star data structures, we must (i) remove  $a$  from every  $\mathcal{D}_{\text{big}}$ , (ii) remove  $a$  from  $\mathcal{D}_{\text{small}}$ . If  $a \in A_{\text{big}}$ , we also (iii) deactivate  $\mathcal{D}_{\text{big}}(a)$ . If  $a \in A_{\text{small}}$ , we also (iv) remove the points of  $\Sigma_a$  from  $\mathcal{D}_{\text{small}}$ . The operations (i-iii) can be performed in  $O(\text{polylog } n)$  time each, but (iv) may take up to  $O(\sqrt{n} \text{ polylog } n)$  time.

On the other hand, there are now  $O(\sqrt{n})$  data structures to query during each relaxation step, as there are  $O(n/\sqrt{n})$  data structures  $\mathcal{D}_{\text{big}}(\cdot)$ . Thus, the query time within each relaxation step is  $O(\sqrt{n} \log^2 n)$ . We can now bound the time spent within the Hungarian search.

► **Lemma 5.3.** *Hungarian search takes  $O(r\sqrt{n} \text{polylog } n)$  time.*

**Updating support stars.** As the flow support changes, the membership of support stars may shift and a big star may eventually become small (or vice versa). To efficiently support this, introduce some fuzziness to when a star should be big or small. Standard charging argument shows that the amortized update time is  $O(r\sqrt{n}(r + \sqrt{n}) \text{polylog } n)$ .

Membership of support stars can only be changed by augmentations, so the number of star membership changes by a single augmenting path is bounded above by twice its length (each vertex may be removed from one star, and/or added to another star). Thus, individual membership changes can be performed in  $O(\text{polylog } n)$  time each, and there are  $O(rn \log n)$  total.

**Preprocessing time.** To build the very first set of data structures, we take  $O(rn \text{polylog } n)$  time. There are  $r|\Sigma_a|$  points in each  $\mathcal{D}_{\text{big}}(a)$ , but the  $\Sigma_a$  are disjoint, so the total points to insert is  $O(rn)$ .  $\mathcal{D}_{\text{small}}$  also has at most  $O(rn)$  points. Each BCP data structure can be constructed in  $O(\text{polylog } n)$  times its size, so the total preprocessing time is  $O(rn \text{polylog } n)$ .

**Between searches.** After an augmentation, we reset the above data structures to their initial state plus the change from the augmentation using the rewinding mechanism. By reversing the sequence of insertions/deletions to each data structure over the course of the Hungarian search, we can recover the versions data structures as they were when the Hungarian search began. This takes time proportional to the time of the Hungarian search,  $O(r\sqrt{n} \text{polylog } n)$  by Lemma 5.3. The most recent augmentation may have deactivated at most one active excess and at most one active deficit, which we can update in the data structures in  $O(\sqrt{n} \text{polylog } n)$  time. Additionally, the augmentation may have changed the membership of some support stars, but we analyzed the time for membership changes earlier. Finally, we note that an augmenting path cannot reduce the support degree of a vertex to zero, and therefore no new dead vertices are created by augmentation.

**Between excess scales.** When the excess scale changes, vertices that were previously inactive may become active, and vertices that were dead may be revived (however, no active vertices deactivate, and no live vertices die as the result of  $\Delta$  decreasing). If we have the data structures built on the active excesses at the end of the previous scale, then we can add in each newly active  $a \in A$  and charge this insertion to the (future) augmenting path or contraction which eventually makes the vertex inactive, or absorbs it into another component. By Theorem 5.1, there are  $O(n \log n)$  such newly active vertices. The time to perform data structure updates for each of them is  $O(\sqrt{n} \text{polylog } n)$ , so the total time spent bookkeeping newly active vertices is  $O(n^{3/2} \text{polylog } n)$ .

**Putting it together.** After  $O(rn \text{polylog } n)$  preprocessing, we spend  $O(r\sqrt{n} \text{polylog } n)$  time each Hungarian search by Lemma 5.3. After each augmentation, we spend the same amount of time (plus  $O(\text{polylog } n)$  extra) to initialize data structures for the next Hungarian search. We spend up to  $O((rn + r^2\sqrt{n}) \text{polylog } n)$  total time making big-small star switching updates. We spend  $O(n^{3/2} \text{polylog } n)$  time activating and reviving vertices. Thus, the algorithm takes



590  $O(rn(r/\sqrt{n} + \sqrt{n}) \text{polylog } n)$  time to produce optimal potentials  $\pi^*$ , from which we can  
 591 recover  $f^*$  in  $O(r\sqrt{n} \text{polylog } n)$  additional time. This completes the proof of Theorem 1.3.

592 **Acknowledgment.** We thank Haim Kaplan for useful discussion and suggesting to use  
 593 Goldberg *et al.* [4] for our approximation algorithm.

## 594 — References —

- 595 1 Pankaj K. Agarwal, Kyle Fox, Debmalya Panigrahi, Kasturi R. Varadarajan, and Allen  
 596 Xiao. Faster algorithms for the geometric transportation problem. *33rd International  
 597 Symposium on Computational Geometry, SoCG 2017, July 4-7, 2017, Brisbane, Australia,*  
 598 7:1–7:16, 2017. [⟨https://doi.org/10.4230/LIPIcs.SocG.2017.7⟩](https://doi.org/10.4230/LIPIcs.SocG.2017.7).
- 599 2 Jack Edmonds and Richard M. Karp. Theoretical improvements in algorithmic efficiency  
 600 for network flow problems. *J. ACM* 19(2):248–264, 1972. [⟨https://doi.org/10.1145/  
 601 321694.321699⟩](https://doi.org/10.1145/321694.321699).
- 602 3 Jeff Erickson. Maximum flows and parametric shortest paths in planar graphs. *Proceedings  
 603 of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010,  
 604 Austin, Texas, USA, January 17-19, 2010*, 794–804, 2010. [⟨https://doi.org/10.1137/1.  
 605 9781611973075.65⟩](https://doi.org/10.1137/1.9781611973075.65).
- 606 4 Andrew V. Goldberg, Sagi Hed, Haim Kaplan, and Robert E. Tarjan. Minimum-cost  
 607 flows in unit-capacity networks. *Theory Comput. Syst.* 61(4):987–1010, 2017. [⟨https://doi.org/10.1007/s00224-017-9776-7⟩](https://doi.org/10.1007/s00224-017-9776-7).
- 608 5 Andrew V. Goldberg and Robert E. Tarjan. Finding minimum-cost circulations by success-  
 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630

## A Proofs from Section 3

### A.1 Preliminaries on Network Flows

**Network.** Let  $G = (V, E)$  be a directed graph, augmented by edge costs  $c$  and capacities  $u$ , and a supply-demand function  $\phi$  defined on the vertices. One can turn the graph  $G$  into a *network*  $N = (V, \vec{E})$ : For each directed edge  $(v, w)$  in  $E$ , insert two *arcs*  $v \rightarrow w$  and  $w \rightarrow v$  into the arc set  $\vec{E}$ ; the *forward arc*  $v \rightarrow w$  inherits the capacity and cost from the directed graph  $G$ , while the *backward arc*  $w \rightarrow v$  satisfies  $u(w \rightarrow v) = 0$  and  $c(w \rightarrow v) = -c(v \rightarrow w)$ . This we ensure that the graph  $(V, \vec{E})$  is *symmetric* and the cost function  $c$  is *antisymmetric* on  $N$ . The positive values of  $\phi(v)$  are referred to as *supply*, and the negative values of  $\phi(v)$  as *demand*. We assume that all capacities are nonnegative, all supplies and demands are integers, and the sum of supplies and demands is equal to zero. A *unit-capacity* network has all its edge capacities equal to 1. In this section we assume all networks are of unit-capacity.

**Pseudoflows.** Given a network  $N := (V, \vec{E}, c, u, \phi)$ , a *pseudoflow* (or *flow* to be short)  $f: \vec{E} \rightarrow \mathbb{Z}^2$  on  $N$  is an antisymmetric function on the arcs of  $N$  satisfying  $f(v \rightarrow w) \leq u(v \rightarrow w)$  for every arc  $v \rightarrow w$ . We sometimes abuse the terminology by allowing pseudoflow to be defined on a directed graph, in which case we are actually referring to the pseudoflow on the corresponding network by extending the flow values antisymmetrically to the arcs. We say that  $f$  *saturates* an arc  $v \rightarrow w$  if  $f(v \rightarrow w) = u(v \rightarrow w)$ ; an arc  $v \rightarrow w$  is *residual* if  $f(v \rightarrow w) < u(v \rightarrow w)$ . The *support* of  $f$  in  $N$ , denoted as  $\text{supp}(f)$ , is the set of arcs with positive flows:

$$\text{supp}(f) := \{v \rightarrow w \in \vec{E} \mid f(v \rightarrow w) > 0\}.$$

Given a pseudoflow  $f$ , we define the *imbalance* of a vertex (with respect to  $f$ ) to be

$$\phi_f(v) := \phi(v) + \sum_{w \rightarrow v \in \vec{E}} f(w \rightarrow v) - \sum_{v \rightarrow w \in \vec{E}} f(v \rightarrow w).$$

We call positive imbalance *excess* and negative imbalance *deficit*; and vertices with positive and negative imbalance *excess vertices* and *deficit vertices*, respectively. A vertex is *balanced* if it has zero imbalance. If all vertices are balanced, the pseudoflow is a *circulation*. The *cost* of a pseudoflow is defined to be

$$\text{cost}(f) := \sum_{v \rightarrow w \in \text{supp}(f)} c(v \rightarrow w) \cdot f(v \rightarrow w).$$

The *minimum-cost flow problem (MCF)* asks to find a circulation of minimum cost inside a given directed graph.

**Residual graph.** Given a pseudoflow  $f$ , one can define the *residual network* as follows. Recall that the set of *residual arcs*  $\vec{E}_f$  under  $f$  are those arcs  $v \rightarrow w$  satisfying  $f(v \rightarrow w) < u(v \rightarrow w)$ . In other words, an arc that is not saturated by  $f$  is a residual arc; similarly, given an arc  $v \rightarrow w$  with positive flow value, the backward arc  $w \rightarrow v$  is a residual arc.

Let  $N = (V, \vec{E}, c, u, \phi)$  be a network constructed from graph  $G$ , with a pseudoflow  $f$  on  $N$ . The *residual graph*  $G_f$  of  $f$  has  $V$  as its vertex set and  $\vec{E}_f$  as its arc set. The *residual capacity*

<sup>2</sup> In general the pseudoflows are allowed to take real-values. Here under the unit-capacity assumption any optimal flows are integer-valued. [«cite integrality theorem?»](#)

668  $u_f$  with respect to pseudoflow  $f$  is defined to be  $u_f(v \rightarrow w) := u(v \rightarrow w) - f(v \rightarrow w)$ . Observe  
 669 that the residual capacity is always nonnegative. We can define residual arcs differently using  
 670 residual capacities:

$$671 \quad \vec{E}_f = \{v \rightarrow w \mid u_f(v \rightarrow w) > 0\}.$$

672 In other words, the set of residual arcs are precisely those arcs in the residual graph, each of  
 673 which has nonzero residual capacity.

674 ► **Lemma 3.1.** *Let  $f$  be an  $\varepsilon$ -optimal pseudoflow in  $G$  and let  $f'$  be an admissible flow in*  
 675  *$G_f$ . Then  $f + f'$  is also  $\varepsilon$ -optimal. ◀◀Lemma 5.3 in [GT90]? Also, where is this used?◀◀*

676 **Proof.** Augmentation by  $f'$  will not change the potentials, so any previously  $\varepsilon$ -optimal arcs  
 677 remain  $\varepsilon$ -optimal. However, it may introduce new arcs  $v \rightarrow w$  with  $u_{f+f'}(v \rightarrow w) > 0$ , that  
 678 previously had  $u_f(v \rightarrow w) = 0$ . We will verify that these arcs satisfy the  $\varepsilon$ -optimality condition.

679 If an arc  $v \rightarrow w$  is newly introduced this way, then by definition of residual capacities  
 680  $f(v \rightarrow w) = u(v \rightarrow w)$ . At the same time,  $u_{f+f'}(v \rightarrow w) > 0$  implies that  $(f + f')(v \rightarrow w) < u(v \rightarrow w)$ .  
 681 This means that  $f'$  augmented flow in the reverse direction of  $v \rightarrow w$  ( $f'(w \rightarrow v) > 0$ ). By  
 682 assumption, the arcs of  $\text{supp}(f')$  are admissible, so  $w \rightarrow v$  was an admissible arc ( $c_\pi(w \rightarrow v) \leq 0$ ).  
 683 By antisymmetry of reduced costs, this implies  $c_\pi(v \rightarrow w) \geq 0 \geq -\varepsilon$ . Therefore, all arcs with  
 684  $u_{f+f'}(v, w) > 0$  respect the  $\varepsilon$ -optimality condition, and thus  $f + f'$  is  $\varepsilon$ -optimal. ◀

685 ► **Lemma 3.3.** *The size of  $\text{supp}(f)$  is at most  $3k$  for any integer circulation  $f$  in reduction*  
 686 *network  $N_H$ . As a corollary, the number of residual backward arcs is at most  $3k$ .*

687 **Proof.** Because  $f$  is a circulation,  $\text{supp}(f)$  can be decomposed into  $k$  paths from  $s$  to  $t$ . Each  
 688  $s$ -to- $t$  path in  $N_H$  is of length three, so the size of  $\text{supp}(f)$  is at most  $3k$ . As every backward  
 689 arc in the residual network must be induced by positive flow in the opposite direction, the  
 690 total number of residual backward arcs is at most  $3k$ . ◀

691 ► **Lemma 3.4.** *Let  $f$  be an  $\varepsilon$ -optimal integer circulation in  $N_H$ , and  $f^*$  be an optimal integer*  
 692 *circulation for  $N_H$ . Then,  $\text{cost}(f) \leq \text{cost}(f^*) + 6k\varepsilon$ .*

693 **Proof.** By Lemma 3.3, the total number of backward arcs in the residual network  $N_f$  is at  
 694 most  $3k$ . Consider the residual flow in  $N_f$  defined by the difference between  $f^*$  and  $f$ . Since  
 695 both  $f$  and  $f^*$  are both circulations and  $N_H$  has unit-capacity, the flow  $f - f^*$  is comprised  
 696 of unit flows on a collection of edge-disjoint residual cycles  $\Gamma_1, \dots, \Gamma_\ell$ . Observe that each  
 697 residual cycle  $\Gamma_i$  must have exactly half of its arcs being backward arcs, and thus we have  
 698  $\sum_i |\Gamma_i| \leq 6k$ .

699 Let  $\pi$  be some potential certifying that  $f$  is  $\varepsilon$ -optimal. Because  $\Gamma_i$  is a residual cycle, we  
 700 have  $c_\pi(\Gamma_i) = c(\Gamma_i)$  since the potential terms telescope. We then see that

$$701 \quad \text{cost}(f) - \text{cost}(f^*) = \sum_i c(\Gamma_i) = \sum_i c_\pi(\Gamma_i) \geq \sum_i (-\varepsilon) \cdot |\Gamma_i| \geq -6k\varepsilon,$$

702 where the second-to-last inequality follows from the  $\varepsilon$ -optimality of  $f$  with respect to  $\pi$ .  
 703 Rearranging the terms we have that  $\text{cost}(f) \leq \text{cost}(f^*) + 6k\varepsilon$ . ◀

## 704 A.2 Multiplicative approximation

705 Let  $T$  be the minimum spanning tree on input graph  $G$  and order its edges by increasing  
 706 length as  $e_1, \dots, e_{r+n-1}$ . Let  $T_\ell$  denote the subgraph of  $T$  obtained by removing the heaviest  
 707  $\ell$  edges in  $T$ . Let  $i$  be the largest index so that the optimal solution to the MPM problem has  
 708 edges between components of  $T_i$ . Choose  $j$  to be the smallest index larger than  $i$  satisfying

709  $c(e_j) \geq kn \cdot c(e_i)$ . For each component  $K$  of  $T_j$ , let  $G_K$  be the subgraph of  $G$  induced on  
 710 vertices of  $K$ ; let  $A_K := K \cap A$  and  $B_K := K \cap B$ , respectively. We partition  $A$  and  $B$  into  
 711 the collection of sets  $A_K$  and  $B_K$  according to the components  $K$  of  $T_j$ . Since  $j < i$ , the  
 712 optimal partial matching in  $G$  can be partitioned into edges between  $A_K$  and  $B_K$  within  
 713  $G_K$ ; no optimal matching edges lie between components.

714 ► **Lemma A.1 (Sharathkumar and Agarwal [SA12, §3.5]).** *Let  $G = (A, B, E_0)$  be the input*  
 715 *to MPM problem, and consider the partitions  $A_K$  and  $B_K$  defined as above. Let  $M^*$  be the*  
 716 *optimal partial matching in  $G$ . Then,*

- 717 (i)  $c(e_i) \leq \text{cost}(M^*) \leq kn \cdot c(e_i)$ , and
- 718 (ii) *the diameter of  $G_K$  is at most  $kn^2 \cdot c(e_i)$  for every  $K \in T_j$ ,*

719 To prove Lemma 3.2, we need to further modify the point set so that the cost of the  
 720 optimal solution does not change, while the diameter of the *whole* point set is bounded. Move  
 721 the points within each component in *translation* so that the minimum distances between  
 722 points across components are at least  $kn \cdot c(e_i)$  but at most  $O(n \cdot kn^2 \cdot c(e_i))$ . This will  
 723 guarantee that the optimal solution still uses edges within the components by Lemma A.1.  
 724 The simplest way of achieving this is by aligning the components one by one into a “straight  
 725 line”, so that the distance between the two farthest components is at most  $O(n)$  times the  
 726 maximum diameter of the cluster.

727 Now one can prove Lemma 3.2 by computing an  $(\varepsilon c(e_i)/6k)$ -optimal circulation  $f$  on the  
 728 point set after translations using additive approximation from Lemma 3.4, together with the  
 729 bound  $c(e_i) \leq \text{cost}(M^*)$  from Lemma A.1.

730 One small problem remains: We need to show that such reduction can be performed in  
 731  $O(n \text{ polylog } n)$  time. Sharathkumar and Agarwal [SA12] have shown that the partition of  $A$   
 732 and  $B$  into  $A_K$ s and  $B_K$ s can be computed in  $O(n \text{ polylog } n)$  time, assuming that the indices  
 733  $i$  and  $j$  can be determined in such time as well. However in our application the choice of  
 734 index  $i$  depends on the optimal solution of MPM problem which we do not know.

735 To solve this issue we perform a binary search on the edges  $e_1, \dots, e_{r+n-1}$ . **«Hmm, we**  
 736 **have no way to check Lemma 4.5(i); but in fact a polynomial bound is good enough.»**  
 737 **«UNRESOLVED ISSUE»**

### 738 A.3 Number of iterations during refinement.

739 To this end we need a bound on the size of the support of  $f$  right before and throughout the  
 740 execution of REFINE. This bound will also be used in the analysis for the running time of  
 741 REFINE.

742 ► **Lemma A.2.** *Let  $f$  be an integer pseudoflow in  $N_H$  with  $O(k)$  excess. Then, the size of*  
 743 *the support of  $f$  is at most  $O(k)$ .*

744 **Proof.** Observe that the reduction graph  $H$  is a directed acyclic graph, and thus the support  
 745 of  $f$  does not contain a cycle. Now  $\text{supp}(f)$  can be decomposed into a set of inclusion-maximal  
 746 paths, each of which contributes a single unit of excess to the flow if the path does not  
 747 terminate at  $t$  or if more than  $k$  paths terminate at  $t$ . By assumption, there are  $O(k)$  units  
 748 of excess to which we can associate to the paths, and at most  $k$  paths (those that terminate  
 749 at  $t$ ) that we cannot associate with a unit of excess. The length of any such paths is at most  
 750 three by construction of the reduction graph  $H$ . Therefore we can conclude that the number  
 751 of arcs in the support of  $f$  is  $O(k)$ . ◀

752 ► **Corollary A.3.** *The size of  $\text{supp}(f)$  is at most  $O(k)$  for pseudoflow  $f$  right before or during*  
 753 *the execution of REFINE.*

754 ► **Lemma 3.5.** *Let  $f$  be a pseudoflow in  $N_H$  with  $O(k)$  excess. The procedure REFINE runs*  
 755 *for  $O(\sqrt{k})$  iterations before the excess of  $f$  becomes zero.*

756 **Proof.** Let  $f_0$  and  $\pi_0$  be the flow and potential at the start of the procedure REFINE. Let  $f$   
 757 and  $\pi$  be the current flow and the potential. Let  $d(v)$  defined to be the amount of potential  
 758 increase at  $v$ , measured in units of  $\varepsilon$ ; in other words,  $d(v) := (\pi(v) - \pi_0(v))/\varepsilon$ .

759 Now divide the iterations executed by the procedure REFINE into two phases: The  
 760 transition from the first phase to the second happens when every excess vertex  $v$  has  
 761  $d(v) \geq \sqrt{k}$ . At most  $\sqrt{k}$  iterations belong to the first phase as each Hungarian search  
 762 increases the potential  $\pi$  by at least  $\varepsilon$  for each excess vertex (and thus increases  $d(v)$  by at  
 763 least one).

764 The number of iterations belonging to the second phase is upper bounded by the amount  
 765 of total excess at the end of the first phase, because each subsequent push of a blocking  
 766 flow reduces the total excess by at least one. We now show that the amount of such excess  
 767 is at most  $O(\sqrt{k})$ . Consider the set of arcs  $E^+ := \{v \rightarrow w \mid f(v \rightarrow w) < f_0(v \rightarrow w)\}$ . The total  
 768 amount of excess is upper bounded by the number of arcs in  $E^+$  that crosses an arbitrarily  
 769 given cut  $X$  that separates the excess vertices from the deficit vertices, when the network  
 770 has unit-capacity [GHKT17, Lemma 3.6]. Consider the set of cuts  $X_i := \{v \mid d(v) > i\}$  for  
 771  $0 \leq i < \sqrt{k}$ ; every such cut separates the excess vertices from the deficit vertices at the end  
 772 of first phase. Each arc in  $E^+$  crosses at most 3 cuts of type  $X_i$  [GHKT17, Lemma 3.1]. So  
 773 there is one  $X_i$  crossed by at most  $3|E^+|/\sqrt{k}$  arcs in  $E^+$ . The size of  $E^+$  is bounded by the  
 774 sum of support sizes of  $f$  and  $f_0$ ; by Corollary A.3 the size of  $E^+$  is  $O(k)$ . This implies an  
 775  $O(\sqrt{k})$  bound on the total excess after the first phase, which in turn bounds the number of  
 776 iterations in the second phase. ◀

## 777 B Proofs from Section 4

778 ► **Lemma 4.1.** *Consider  $\tilde{\pi}$  an  $\varepsilon$ -optimal potential on  $\tilde{H}_f$  and  $\pi$  the corresponding potential*  
 779 *constructed on  $H_f$ . Then,*

- 780 1. *potential  $\pi$  is  $\varepsilon$ -optimal on  $H_f$ , and*
- 781 2. *if arc  $\text{short}(\Pi)$  is admissible under  $\tilde{\pi}$ , then every arc in  $\Pi$  is admissible under  $\pi$ .*

782 **Proof.** Reduced costs for any arc from a normal vertex another is unchanged under either  $\tilde{\pi}$   
 783 or  $\pi$ . Recall that a null path is comprised of one  $A$ -to- $B$  arc, and one or two zero-cost arcs  
 784 (connecting the null vertex/vertices to  $s$  and/or  $t$ ). With our choice of null vertex potentials,  
 785 we observe that the zero-cost arcs still have zero reduced cost. It remains to prove that an  
 786 arbitrary **«residual?»** arc  $(a, b)$  **«arc or directed edge?»** satisfies the  $\varepsilon$ -optimality condition  
 787 and admissibility when either  $a$  or  $b$  is a null vertex.

788 By construction of the shortcut graph, there is always a null path  $\Pi$  that contains  $(a, b)$ .  
 789 Observe that  $c_\pi(a, b) = c_\pi(\Pi)$ , independent to the type of null path. Again by construction,  
 790  $c_\pi(\Pi) = c_{\tilde{\pi}}(\text{short}(\Pi))$ , so we have  $c_\pi(a, b) = c_{\tilde{\pi}}(\text{short}(\Pi)) \geq -\varepsilon$ . Additionally, if  $\text{short}(\Pi)$  is  
 791 admissible under  $\tilde{\pi}$ , then so is  $(a, b)$  under  $\pi$ . This proves the lemma. ◀

### 792 B.1 Dynamic data structures for search procedures

793 Here we formally describe in details the set of dynamic data structure we use for the  
 794 Hungarian search and depth-first search procedures.

795 For Hungarian search, we maintain the following for each type of outgoing arcs of  $\tilde{H}_f$   
 796 leaving  $\tilde{S}$ :

- 797 1. Non-shortcut backward arcs  $(v, w)$  with  $(w, v) \in \text{supp}(f)$ . For these, we can maintain a  
798 min-heap on  $\text{supp}(f)$  arcs as each  $v$  arrives in  $\tilde{S}$ .
- 799 2. Non-shortcut  $A$ -to- $B$  forward arcs. For these, we can use a BCP data structure between  
800  $(A \setminus A_\emptyset) \cap \tilde{S}$  and  $(B \setminus B_\emptyset) \setminus \tilde{S}$ , weighted by potential.
- 801 3. Non-shortcut forward arcs from  $s$ -to- $A$  and from  $B$ -to- $t$ . For  $s$ , we can maintain a  
802 min-heap on the potentials of  $B \setminus \tilde{S}$ , queried while  $s \in \tilde{S}$ . For  $t$ , we can maintain a  
803 max-heap on the potentials of  $A \cap \tilde{S}$ , queried while  $t \notin \tilde{S}$ .
- 804 4. Shortcut arcs  $(s, b)$  corresponding to null 2-paths from  $s$  to  $b \in (B \setminus B_\emptyset) \setminus S$ . For these, we  
805 maintain a BCP data structure with  $P = A_\emptyset$ ,  $Q = (B \setminus B_\emptyset) \setminus S$  with weights  $\omega(p) = \pi(s)$   
806 for all  $p \in P$ , and  $\omega(q) = \pi(q)$  for all  $q \in Q$ . A response  $(a, b)$  corresponds to th null  
807 2-path  $(s, a, b)$ . This is only queried while  $s \in S$ .
- 808 5. Shortcut arcs  $(a, t)$  corresponding to null 2-paths from  $a \in (A \setminus A_\emptyset) \cap S$  to  $t$ . For these,  
809 we maintain a BCP data structure with  $P = (A \setminus A_\emptyset) \cap S$ ,  $Q = B_\emptyset \setminus S$  with weights  
810  $\omega(p) = \pi(p)$  for all  $p \in P$ , and  $\omega(q) = \pi(t)$  for all  $q \in Q$ . A response  $(a, b)$  corresponds to  
811 th null 2-path  $(a, b, t)$ . This is only queried while  $t \notin S$ .
- 812 6. Shortcut arcs  $(s, t)$  corresponding to null 3-paths. For these, we maintain in a BCP data  
813 structure with  $P = A_\emptyset \setminus S$ ,  $Q = B_\emptyset \setminus S$  with weights  $\omega(p) = \pi(s)$  for all  $p \in P$ , and  
814  $\omega(q) = \pi(t)$  for all  $q \in Q$ . A response  $(a, b)$  corresponds to th null 3-path  $(s, a, b, t)$ . This  
815 is only queried while  $s \in S$  and  $t \notin S$ .

816 For depth-first search, we maintain the following for each type of outgoing arcs of  $\tilde{H}_f$   
817 leaving  $\tilde{S}$ :

- 818 1. Non-shortcut backward arcs  $(v', w')$  with  $(w', v') \in \text{supp}(f)$ . For these, we can maintain  
819 a min-heap on  $(w', v') \in \text{supp}(f)$  arcs for each normal  $v' \in V$ .
- 820 2. Non-shortcut  $A$ -to- $B$  forward arcs. For these, we maintain a NN data structure over  
821  $P = (B \setminus B_\emptyset) \setminus \tilde{S}$ , with weights  $\omega(p) = \pi(p)$  for each  $p \in P$ . We subtract  $\pi(v')$  from the  
822 NN distance to recover the reduced cost of the arc from  $v'$ .
- 823 3. Non-shortcut forward arcs from  $s$ -to- $A$  and from  $B$ -to- $t$ . For  $s$ , we can maintain a  
824 min-heap on the potentials of  $B \setminus \tilde{S}$ , queried only if  $v' = s$ . For  $B$ -to- $t$  arcs, there is only  
825 one arc to check if  $v' \in B$ , which we can examine manually.
- 826 4. Shortcut arcs  $(s, b)$  corresponding to null 2-paths from  $s$  to  $b \in (B \setminus B_\emptyset) \setminus S$ . For these, we  
827 maintain a NN data structure with  $P = A_\emptyset$ ,  $Q = (B \setminus B_\emptyset) \setminus S$  with weights  $\omega(p) = \pi(s)$   
828 for all  $p \in P$ , and  $\omega(q) = \pi(q)$  for all  $q \in Q$ . A response  $(a, b)$  corresponds to th null  
829 2-path  $(s, a, b)$ . This is only queried if  $v' = s$ .
- 830 5. Shortcut arcs  $(a, t)$  corresponding to null 2-paths from  $a \in (A \setminus A_\emptyset) \cap S$  to  $t$ . For these,  
831 we maintain a NN data structure over  $P = B_\emptyset \setminus S$  with weights  $\omega(p) = \pi(t)$  for each  
832  $p \in P$ . A response  $(v', b)$  corresponds to th null 2-path  $(v', b, t)$ . We subtract  $\pi(v')$  from  
833 the NN distance to recover the reduced cost of the arc from  $v'$ . This is not queried if  
834  $t \in \tilde{S}$ .
- 835 6. Shortcut arcs  $(s, t)$  corresponding to null 3-paths. For these, we maintain in a NN data  
836 structure with  $P = A_\emptyset \setminus S$ ,  $Q = B_\emptyset \setminus S$  with weights  $\omega(p) = \pi(s)$  for all  $p \in P$ , and  
837  $\omega(q) = \pi(t)$  for all  $q \in Q$ . A response  $(a, b)$  corresponds to th null 3-path  $(s, a, b, t)$ . This  
838 is only queried while  $v' = s$  and  $t \notin S$ .

## 839 B.2 Number of relaxations

840 First we bound the number of relaxations performed by both the Hungarian search and the  
841 depth-first search.

842 ► **Lemma B.1.** *Hungarian search performs  $O(k)$  relaxations before a deficit vertex is reached.*



**Proof.** **«TO BE REWRITTEN.»** First we prove that there are  $O(k)$  non-shortcut relaxations. Each edge relaxation adds a new vertex to  $S$ , and non-shortcut relaxations only add normal vertices. The vertices of  $V \setminus S$  fall into several categories: (i)  $s$  or  $t$ , (ii) vertices of  $A$  or  $B$  with 0 imbalance, and (iii) deficit vertices of  $A$  or  $B$  ( $S$  contains all excess vertices). The number of vertices in (i) and (iii) is  $O(k)$ , leaving us to bound the number of (ii) vertices.

An  $A$  or  $B$  vertex with 0 imbalance must have an even number of  $\text{supp}(f)$  edges. There is either only one positive-capacity incoming arc (for  $A$ ) or outgoing arc (for  $B$ ), so this quantity is either 0 or 2. Since the vertex is normal, this must be 2. We charge 0.5 to each of the two  $\text{supp}(f)$  arcs; the arcs of  $\text{supp}(f)$  have no more than 1 charge each. Thus, the number of type (ii) vertex relaxations is  $O(|\text{supp}(f)|)$ . By Corollary A.3,  $O(|\text{supp}(f)|) = O(k)$ .

Next we prove that there are  $O(k)$  shortcut relaxations. Recall the categories of shortcuts from the list of data structures above. We have shortcuts corresponding to (i) null 2-paths surrounding  $a \in A_\emptyset$ , (ii) null 2-paths surrounding  $b \in B_\emptyset$ , and (iii) null 3-paths, which go from  $s$  to  $t$ . There is only one relaxation of type (iii), since  $t$  can only be added to  $S$  once. The same argument holds for type (ii).

Each type (i) relaxation adds some normal  $b \in B \setminus B_\emptyset$  into  $S$ . Since  $b$  is normal, it must either have deficit or an adjacent arc of  $\text{supp}(f)$ . We charge this relaxation to  $b$  if it is deficit, or the adjacent arc of  $\text{supp}(f)$  otherwise. No vertex is charged more than once, and no  $\text{supp}(f)$  edge is charged more than twice, therefore the total number of type (i) relaxations is  $O(|\text{supp}(f)|)$ . By Corollary A.3,  $O(|\text{supp}(f)|) = O(k)$ . ◀

Similarly we can prove that there are  $O(k)$  relaxations during the DFS.

► **Corollary B.2.** *Depth-first search performs  $O(k)$  relaxations before a deficit vertex is reached.*

### B.3 Time analysis

Now we complete the time analysis by showing that each Hungarian search and depth-first search can be implemented in  $O(k \text{ polylog } n)$  time after a one-time  $O(n \text{ polylog } n)$ -time preprocessing.

► **Lemma B.3.** *After  $O(n \text{ polylog } n)$ -time preprocessing, each Hungarian search can be implemented in  $O(k \text{ polylog } n)$  time.*

**Proof.** Each of the constant number of data structures used by the Hungarian search can be constructed in  $O(n \text{ polylog } n)$  time. For each data structure queried during a relaxation, the new vertex moved into  $S$  causes a constant number of updates, each of which can be implemented in  $O(\text{polylog } n)$  time. We first prove that the number of BCP operations during the Hungarian search over is bounded by  $O(k)$ .

1. Let  $S^t$  denote the initial set  $S$  at the beginning of the  $t$ -th Hungarian search, Assume for now that, at the beginning of the  $(t + 1)$ -th Hungarian search, we have the set  $S^t$  from the previous iteration. To construct  $S^{t+1}$ , we remove the vertices that had excess decreased to zero by the  $t$ -th blocking flow. Thus, we are able to initialize  $S$  at the cost of one BCP deletion per excess vertex, which sums to  $O(k)$  over the entire course of REFIN. **«Too strong as a bound? Is it enough to look at one Hungarian search?»**
2. During each Hungarian search, a vertex entering  $S$  may incur one BCP insertion/deletion. We can charge the updates to the number of relaxations over the course of Hungarian search. The number of relaxations in a Hungarian search is  $O(k)$  by Lemma B.1.

886 3. To obtain  $S^t$ , we keep track of the points added to  $S^t$  since the last Hungarian search.  
 887 After the augmentation, we remove those points added to  $S^t$ . By (2) there are  $O(k)$  such  
 888 points to be deleted, so reconstructing  $S^t$  takes  $O(k)$  BCP operations.

889 For potential updates, we use the same trick by Vaidya [Vai89] to lazily update potentials  
 890 after vertices leave  $S$  (similar to Lemma ??), but this time only for normal vertices. Normal  
 891 vertices are stored in each data structure with weight  $\omega(v) = \pi(v) - \delta$ , and  $\delta$  is increased in  
 892 lieu of increasing the potential of vertices in  $S$ . When a vertex leave  $S$  (through the rewind  
 893 mechanism above), we restore its potential as  $\pi(v) \leftarrow \omega(v) + \delta$ . With lazy updates, the  
 894 number of potential updates on normal vertices is bounded by the number of relaxations in  
 895 the Hungarian search, which is  $O(k)$  by Lemma B.1. Note that null vertex potentials are not  
 896 handled in the Hungarian search. **«then where? Lemma B.6»** ◀

897 There are no potentials to update within DFS, so the running time of DFS boils down  
 898 to the time spent to querying and updating the data structures.

899 ► **Lemma B.4.** *After  $O(n \text{ polylog } n)$ -time preprocessing, each depth-first search can be*  
 900 *implemented in  $O(k \text{ polylog } n)$  time.*

901 **Proof.** At the beginning of REFINe, we can initialize the  $O(1)$  data structures used in DFS  
 902 in  $O(n \text{ polylog } n)$  time. We use the same rewinding mechanism as in Hungarian search  
 903 (Lemma B.3) to avoid reconstructing the data structures across iterations of REFINe, so  
 904 the total time spent is bounded by the  $O(\text{polylog } n)$  times the number of relaxations. By  
 905 Corollary B.2, the running time for depth-first search is  $O(k \text{ polylog } n)$ . ◀

## 906 B.4 Number of potential updates on null vertices

907 In our implementation of REFINe, we do not explicitly construct  $\tilde{H}_f$ ; instead we query its  
 908 edges using BCP/NN oracles and min/max heaps on elements of  $H_f$ . Potentials on the null  
 909 vertices are only required right before an augmentation sends a flow through a null path,  
 910 making the null vertices it passes normal. We use the construction from Lemma 4.1 to obtain  
 911 potential  $\pi$  on  $H_f$  such that the flow  $f$  is both  $\varepsilon$ -optimal and admissible with respect to  $\pi$ .

912 **Size of blocking flows.** Now we bound the total number arcs whose flow is updated by a  
 913 blocking flow during the course of REFINe. This bounds both the time spent updating the  
 914 flow on these arcs and also the time spent on null vertex potential updates (Lemma B.6).

915 ► **Lemma B.5.** *The support of each blocking flow found in REFINe is of size  $O(k)$ .*

916 **Proof.** Let  $i$  be fixed and consider the invocation of DFS which produces the  $i$ -th blocking  
 917 flow  $f_i$ . DFS constructs  $f_i$  as a sequence of admissible excess-deficit paths, which appear as  
 918 path  $P$  in Algorithm ??. Every arc in  $P$  is an arc relaxed by DFS, so  $N_i$  is bounded by the  
 919 number of relaxations performed in DFS. Using Corollary B.2, we have  $N_i = O(k)$ . ◀

920 ► **Lemma B.6.** *The number of end-of-REFINE null vertex potential updates is  $O(n)$ . The*  
 921 *number of augmentation-induced null vertex potential updates in each invocation of REFINe*  
 922 *is  $O(k \log k)$ .*

923 **Proof.** The number of end-of-REFINE potential updates is  $O(n)$ . Each update due to flow  
 924 augmentation involves a blocking flow sending positive flow through a null path, causing a  
 925 potential update on the passed null vertex. We charge this potential update to the edges  
 926 of that null path, which are in turn arcs with positive flow in the blocking flow. For each

blocking flow, no positive arc is charged more than twice. It follows that the number of augmentation-induced updates is at most the size of support of the blocking flow, which is  $O(k)$  by Lemma B.5. According to Lemma 3.5 there are  $O(\sqrt{k})$  iterations of REFINE before it terminates. Summing up we have an  $O(k\sqrt{k})$  bound over the course of REFINE. ◀

Now combining Lemma B.3, Lemma B.4, and Lemma B.6 completes the proof of Theorem 1.2.

## C Proofs from Section 5

### C.1 Recovering the optimal flow

We use the recovery strategy from Agarwal *et al.* [AFP<sup>+</sup>17], which runs in  $O(n \text{ polylog } n)$  time. The main idea is that, if  $\mathcal{T}$  is an undirected *spanning tree of admissible edges* under optimal potentials  $\pi^*$ , then there exists an optimal flow  $f^*$  with support only on arcs corresponding to edges of  $\mathcal{T}$ . Intuitively,  $\mathcal{T}$  is a maximal set of linearly independent dual LP constraints for the optimal dual ( $\pi^*$ ), so there exists an optimal primal solution ( $f^*$ ) with support only on these arcs. To see this, we can use a perturbation argument: raising the cost of each non-tree edge by  $\varepsilon > 0$  does not change  $\text{cost}(\pi^*)$  or the feasibility of  $\pi^*$ , but does raise the cost of any circulation  $f$  using non-tree edges. Strong duality suggests that  $\text{cost}(f^*) = \text{cost}(\pi^*)$  is unchanged, therefore  $f^*$  must have support only on the tree edges.

Since the arcs corresponding to edges of  $\mathcal{T}$  have no cycles, we can solve the maximum flow in linear time using the following greedy algorithm. Let

(v) be the parent of vertex  $v$  in  $\mathcal{T}$ . We begin with  $f^*(v) = 0$  and process  $\mathcal{T}$  from its leaves upwards.

For a supply leaf  $v$ , we satisfy its supply by choosing  $f^*(v) \leftarrow \phi(v)$ . Otherwise if leaf  $v$  is a demand vertex, we choose  $f^*(v) \leftarrow -\phi(v)$ . Once we've solved the supplies/demands for each leaf, then we can trim the leaves, removing them from  $\mathcal{T}$  and setting the supply/demand of each parent-of-a-leaf to its current imbalance. Then, we can recurse on this smaller tree and its new set of leaves.

► **Lemma C.1.** *Let  $G(\mathcal{T})$  be the subnetwork of  $G$  corresponding to edges of the undirected spanning tree  $\mathcal{T}$ . If there exists a flow in  $G(\mathcal{T})$  which satisfies every supply and demand, then the greedy algorithm finds the maximum flow in  $G(\mathcal{T})$  in  $O(n)$  time.*

► **Proof.** Observe that, for any flow  $f$  in  $G$ ,  $\text{supp}(f)$  has no paths of length longer than one. Thus, if a flow  $f^*$  satisfying supplies/demands exists within  $G(\mathcal{T})$ , then each supply vertex has flow paths that terminate at its parent/children. Similarly, each demand vertex receive all its flow from its parent/children. Since there is only one option for a supply leaf (resp. demand leaf) to send its flow (resp. receive its flow), the greedy algorithm correctly identifies the values of  $f^*$  for arcs adjoining  $\mathcal{T}$  leaves. Trimming these leaves, we can apply this argument recursively for their parents. The running time of the greedy algorithm is  $O(n)$ , as leaves can be identified in  $O(n)$  time and no vertex becomes a leaf more than once. ◀

It remains to show how we construct  $\mathcal{T}$ . We begin with a (spanning) *shortest path tree* (SPT)  $T$  in the residual network of  $f$ , under reduced costs and rooted at an arbitrary vertex  $r$ . For the SPT to span, we need the additional assumption that  $G$  is strongly connected. We can make  $G$  strongly connected by adding a 0-supply vertex  $s$  with arcs  $s \rightarrow a$  for all  $a \in A$  and  $b \rightarrow a$  for all  $b \in B$ , with some high cost  $M$ . Following Orlin [Orl93], these arcs cannot appear in an optimal flow if  $M$  is sufficiently high, and we can extend  $\pi^*$  to include  $s$  using  $\pi^*(s) = 0$  if  $M > \max_{b \in B} \pi^*(b)$ . This extension to  $\pi^*$  preserves feasibility.

It remains to show how we construct  $\mathcal{T}$ . We begin with a (spanning) *shortest path tree* (SPT)  $T$  in the residual network of  $f$ , under reduced costs and rooted at an arbitrary vertex  $r$ . For the SPT to span, we need the additional assumption that  $G$  is strongly connected. We can make  $G$  strongly connected by adding a 0-supply vertex  $s$  with arcs  $s \rightarrow a$  for all  $a \in A$  and  $b \rightarrow a$  for all  $b \in B$ , with some high cost  $M$ . Following Orlin [Orl93], these arcs cannot appear in an optimal flow if  $M$  is sufficiently high, and we can extend  $\pi^*$  to include  $s$  using  $\pi^*(s) = 0$  if  $M > \max_{b \in B} \pi^*(b)$ . This extension to  $\pi^*$  preserves feasibility.

965 The edges corresponding to arcs of  $T$  do not suffice for  $\mathcal{T}$ , since some SPT arcs may be  
 966 inadmissible. Let  $d_r(v)$  be the shortest path distance of  $v \in A \cup B \cup \{s\}$  from  $r$ , and consider  
 967 potentials  $\pi = \pi^* - d_r$ .

968 ► **Lemma C.2 Orlin [Ori93] Lemma 3.** *Let  $f$  be a flow satisfying the optimality conditions*  
 969 *with respect to  $\pi^*$ . Then, (i)  $f$  satisfies the optimality conditions with respect to  $\pi$ , and (ii)*  
 970 *all SPT arcs are admissible under  $\pi$ .*

971 We can use this lemma to argue that  $\pi$  is still optimal. Recall that  $f$  has values defined  
 972 only on the non-contracted residual arcs; we can apply the first part of Lemma C.2 on these  
 973 arcs. For arcs within contracted components, we use a different argument. Observe that  
 974 each  $\hat{v} \in \hat{V}$  is spanned by a set of  $\text{supp}(f)$  arcs, which are admissible by invariant. Thus, all  
 975  $v \in \hat{v}$  are equidistant from  $r$ , and they will have the same value  $d_r(v)$ . It follows that the  
 976 reduced costs of arcs with both endpoints in  $\hat{v}$  do not change when replacing  $\pi^*$  with  $\pi$ , so  
 977 arcs contained in  $\hat{v}$  that met the optimality conditions for  $\pi^*$  still meet them for  $\pi$ .

978 From the second part of Lemma C.2, the SPT  $T$  is a spanning tree of admissible arcs  
 979 under  $\pi$ . We set  $\mathcal{T}$  to be the set of undirected edges corresponding to  $T$ .

980 **Computing the SPT.** We conclude by describing the procedure for building the SPT, i.e.  
 981 by running Dijkstra's algorithm in the residual network. We use a geometric implementation  
 982 that is very similar to Hungarian search. We begin with  $S = \{r\}$  and  $d_r(r) = 0$ , where  $r$  is  
 983 our arbitrary root. For all other vertices,  $d_r(v)$  is initially unknown. In each iteration, we  
 984 relax the minimum-reduced cost arc  $v \rightarrow w$  in the frontier  $S \times (A \cup B) \setminus S$ , adding  $w$  to  $S$ , and  
 985 setting  $d_r(w) = d_r(v) + c_{\pi^*}(v, w)$ . Once  $S = A \cup B$ , the SPT  $T$  is the set of relaxed arcs.

986 If an either direction of an arc of  $\text{supp}(f)$  enters the frontier, we relax it immediately.  
 987 To detect support arcs, we build a list for each  $v \in A \cup B$  of the support arcs which use  
 988  $v$  as an endpoint, and once  $v \in S$  we check its list. There are  $O(n)$  support arcs in total  
 989 (by acyclicity of  $E(\text{supp}(f))$ ), so the total time spent searching these lists is  $O(n)$ . Such  
 990 relaxations are correct for the shortest path tree, since the support edges are admissible and  
 991 reduced costs are nonnegative.

992 Other edges appearing in the frontier can be split into three categories:

- 993 1. Forward  $A$ -to- $B$  arcs. We query these using a BCP with  $P = A \cap S$  and  $Q = B \setminus S$ .
- 994 2.  $B$ -to- $s$  arcs. These will never have flow support. We can query the minimum with a  
 995 max-heap on potentials of  $B \cap S$ . We query these while  $s \in S$ .
- 996 3.  $s$ -to- $A$  arcs. These will also never have flow support. We can query the minimum with a  
 997 min-heap on potentials of  $A \setminus S$ . We query these while  $s \in S$ .

998 We perform  $O(n)$  relaxations and takes  $O(\text{polylog } n)$  time per relaxation, for non-support  
 999 relaxations. An additional  $O(n)$  time is spent relaxing support edges. The total running  
 1000 time of Dijkstra's algorithm is  $O(n \text{ polylog } n)$ . Combining with Lemma C.1, we obtain the  
 1001 following.

1002 ► **Lemma C.3.** *Given optimal potentials  $\pi^*$  and an optimal contracted flow  $f$ , the optimal*  
 1003 *flow  $f^*$  can be computed in  $O(n \text{ polylog } n)$  time.*

## 1004 C.2 Recovering the optimal flow for sum-of-distances.

1005 When the matching objective uses the just the  $p$ -norms (that is, when  $q = 1$ ), we can prove  
 1006 that the subgraph formed by admissible arcs is in fact *planar*. Planarity gives us two things  
 1007 towards a simple  $f^*$  recovery: there are only a linear number of admissible arcs, and the  
 1008 max-flow on them can be solved in near-linear time with planar graph max-flow algorithms.

Up until now, we have not placed restrictions on coincidence between  $A$  and  $B$ , but for the next proof it is useful to do so. We can assume that all points within  $A \cup B$  are distinct, otherwise we can replace all points coincident at  $x \in \mathbb{R}^2$  with a single point whose supply/demand is  $\sum_{v \in A \cup B: v=x} \lambda(v)$ . This is roughly equivalent to transporting as much as we can between coincident supply and demand, and is optimal by triangle inequality.

Without loss of generality, assume  $\pi^*$  is nonnegative (raising  $\pi^*$  uniformly on all points does not change the objective or feasibility). Recall that  $\pi^*$  is feasibility if for all  $a \in A$  and  $b \in B$

$$c_{\pi^*}(a, b) = \|a - b\|_p - \pi^*(a) + \pi^*(b) \geq 0.$$

An arc  $a \rightarrow b$  is admissible when

$$c_{\pi^*}(a, b) = \|a - b\|_p - \pi^*(a) + \pi^*(b) = 0.$$

We note that these definitions have a nice visual: Place disks  $D_q$  of radius  $\pi(q)$  at each  $q \in A \cup B$ . Feasibility states that for all  $a \in A$  and  $b \in B$ ,  $D_a$  cannot contain  $D_b$  with a gap between their boundaries. The arc  $a \rightarrow b$  is admissible when  $D_a$  contains  $D_b$  and their boundaries are tangent.

► **Lemma C.4.** *Let  $\pi^*$  be a set of optimal potentials for the point sets  $A$  and  $B$ , under costs  $c(a, b) = \|a - b\|_p$ . Then, the set of admissible arcs under  $\pi^*$  form a planar graph.*

**Proof.** We assume the points of  $A \cup B$  are in general position (e.g. by symbolic perturbation) such that no three points are collinear. Let  $a_1 \rightarrow b_1$  and  $a_2 \rightarrow b_2$  be any pair of admissible arcs under  $\pi^*$ . We will isolate them from the rest of the points, considering  $\pi^*$  restricted to the four points  $\{a_1, a_2, b_1, b_2\}$ . Clearly, this does not change whether the two arcs cross. Observe that we can raise  $\pi^*(a_2)$  and  $\pi^*(b_2)$  uniformly, until  $c_{\pi^*}(a_2, b_1) = 0$ , without breaking feasibility or changing admissibility of  $a_1 \rightarrow b_1$  and  $a_2 \rightarrow b_2$ . Henceforth, we assume that we have modified  $\pi^*$  in this way to make  $a_2 \rightarrow b_1$  admissible. Given positions of  $a_1, a_2$ , and  $b_1$ , we now try to place  $b_2$  such that  $a_1 \rightarrow b_1$  and  $a_2 \rightarrow b_2$  cross. Specifically,  $b_2$  must be placed within a region  $\mathcal{F}$  that lies between the rays  $\overrightarrow{a_2 a_1}$  and  $\overrightarrow{a_2 b_1}$ , and within the halfplane bounded by  $\overleftrightarrow{a_1 b_1}$  that does not contain  $a_2$ .

Let  $g_a(q) := \|a - q\| - \pi^*(a)$  for  $a \in A$  and  $q \in \mathbb{R}^2$ . Let the *bisector* between  $a_1$  and  $a_2$  be  $\beta := \{q \in \mathbb{R}^2 \mid g_{a_1}(q) = g_{a_2}(q)\}$ .  $\beta$  is a curve subdividing the plane into two open faces, one where  $g_{a_1}$  is minimized and the other where  $g_{a_2}$  is. From these definitions, admissibility of  $a_1 \rightarrow b_1$  and  $a_2 \rightarrow b_1$  imply that  $b_1$  is a point of the bisector.

We show that  $\mathcal{F}$  lies entirely on the  $g_{a_1}$  side of the bisector. First, we prove that the closed segment  $\overline{a_1 b_1}$  lies entirely on the  $g_{a_1}$  side, except  $b_1$  which lies on  $\beta$ . Any  $q \in \overline{a_1 b_1}$  can be written parametrically as  $q(t) = (1 - t)b_1 + ta_1$  for  $t \in [0, 1]$ . Consider the single-variable functions  $g_{a_1}(q(t))$  and  $g_{a_2}(q(t))$ .

$$\begin{aligned} g_{a_1}(q(t)) &= (1 - t)\|a_1 - b_1\| - \pi(a_1) \\ g_{a_2}(q(t)) &= \|(a_2 - b_1) - t(a_1 - b_1)\| - \pi(a_2) \end{aligned}$$

At  $t = 0$ , these expressions are equal. Observe that the derivative with respect to  $t$  of  $g_{a_1}(q(t))$  is less than  $g_{a_2}(q(t))$ . Indeed, the value of  $\frac{d}{dt}\|(a_2 - b_1) - t(a_1 - b_1)\|$  is at least  $-\|a_1 - b_1\| = \frac{d}{dt}g_{a_1}(q(t))$ , which is realized if and only if  $\frac{(a_2 - b_1)}{\|a_2 - b_1\|} = \frac{(a_1 - b_1)}{\|a_1 - b_1\|}$ . This corresponds to  $\overrightarrow{a_2 b_1}$  and  $\overrightarrow{a_1 b_1}$  being parallel, but this is disallowed since  $a_1, a_2, b_1$  are in general position. Thus,  $g_{a_1}(q(t)) \leq g_{a_2}(q(t))$  with equality only at  $b_1$ .

Now, we parameterize each point of  $\mathcal{F}$  in terms of points on  $\overline{a_1 b_1}$ . Every  $q \in \mathcal{F}$  can be written as  $q(t') = q' + t'(q' - a_2)$  for some  $q' \in \overline{a_1 b_1}$  and  $t \geq 0$ , i.e.  $q' = \overline{a_1 b_1} \cap \overrightarrow{a_2 q}$ .

We call  $q'$  the *projection* of  $q$  onto  $\overline{a_1 b_1}$ . We can write  $g_{a_1}$  and  $g_{a_2}$  in terms of  $t'$  and observe that  $\frac{d}{dt'} g_{a_1}(q(t')) \leq \frac{d}{dt'} g_{a_2}(q(t'))$ , as the derivative of  $g_a(q(t'))$  is maximized if  $(q(t') - a)$  is parallel to  $(q(t') - a_2)$  and lower otherwise. Notably,  $q(t')$  with projection  $b_1$  have  $\frac{d}{dt'} g_{a_1}(q(t')) < \frac{d}{dt'} g_{a_2}(q(t'))$ , since  $a_1, a_2, b_1$  are in general position. Any  $q(t')$  with a different projection do not have strict inequality, but the projection itself has  $g_{a_1}(q') < g_{a_2}(q')$  for  $q' \neq b_1$  since it lies on  $\overline{a_1 b_1}$ . Therefore, for all  $q \in \mathcal{F} \setminus \{b_1\}$ ,  $g_{a_1}(q') < g_{a_2}(q')$ , and  $\mathcal{F}$  lies on the  $g_{a_1}$  side of the bisector except for  $b_1$  which lies on  $\beta$ . We can eliminate  $b_1$  as a candidate position for  $b_2$ , since points of  $B$  cannot coincide.

Observe that  $g_{a_1}(b) < g_{a_2}(b)$  for  $b \in B$  implies that  $c_\pi(a_1, b) < c_\pi(a_2, b)$ , and  $c_\pi(a_1, b) = c_\pi(a_2, b)$  if and only if  $b$  lies on  $\beta$ . This holds for all  $b \in \mathcal{F}$  including our prospective  $b_2$ , but then  $c_\pi(a_1, b_2) < c_\pi(a_2, b_2) = 0$  since  $a_2 \rightarrow b_2$  is admissible. This violates feasibility of  $a_1 \rightarrow b_2$ , so there is no feasible placement of  $b_2$  which also crosses  $a_1 \rightarrow b_1$  with  $a_2 \rightarrow b_2$ . ◀

We can construct the entire set of admissible arcs by repeatedly querying the minimum-reduced-cost outgoing arc for each  $a \in A$  until the result is not admissible. By Lemma C.4 the resulting arc set forms a planar graph, so by Euler's formula the number of arcs to query is  $O(n)$ . We can then find the maximum flow in time  $O(n \log n)$  time, using for example the planar maximum-flow algorithm by Erickson [Eri10]. *◀cite others like Klein▶*

► **Lemma C.5.** *If the transportation objective is sum-of-costs, then given the optimal potentials  $\pi^*$ , we can compute an optimal flow  $f^*$  in  $O(n \text{ polylog } n)$  time.*

By prioritizing the relaxation of support arcs, we also have the following lemma.

► **Lemma C.6 (Agarwal et al. [AFP<sup>+</sup>17]).** *If arcs of  $\text{supp}(f)$  are relaxed first as they arrive on the frontier, then  $E(\text{supp}(f))$  is acyclic.*

**Proof.** Let  $f_i$  be the pseudoflow after the  $i$ -th augmentation, and let  $T_i$  be the forest of relaxed arcs generated by the Hungarian search for the  $i$ -th augmentation. Namely, the  $i$ -th augmenting path is an excess-deficit path in  $T_i$ , and all arcs of  $T_i$  are admissible by the time the augmentation is performed. Let  $E(T_i)$  be the undirected edges corresponding to arcs of  $T_i$ . Notice that,  $E(\text{supp}(f_{i+1})) \subseteq E(\text{supp}(f_i)) \cup E(T_i)$ . We prove that  $E(\text{supp}(f_i)) \cup E(T_i)$  is acyclic by induction on  $i$ ; as  $E(\text{supp}(f_{i+1}))$  is a subset of these edges, it must also be acyclic. At the beginning with  $f_0 = 0$ ,  $E(\text{supp}(f_0))$  is vacuously acyclic.

Let  $E(\text{supp}(f_i))$  be acyclic by induction hypothesis. Since  $T_i$  is a forest (thus, acyclic), any hypothetical cycle  $\Gamma$  that forms in  $E(\text{supp}(f_i)) \cup E(T_i)$  must contain edges from both  $E(\text{supp}(f_i))$  and  $E(T_i)$ . To give a visual analogy, we will color  $e \in \Gamma$  *purple* if  $e \in E(\text{supp}(f_i)) \cap E(T_i)$ , *red* if  $e \in E(\text{supp}(f_i))$  but  $e \notin E(T_i)$ , and *blue* if  $e \in E(T_i)$  but  $e \notin E(\text{supp}(f_i))$ . Then,  $\Gamma$  is neither entirely red nor entirely blue. We say that red and purple edges are *red-tinted*, and similarly blue and purple edges are *blue-tinted*. Roughly speaking, our implementation of the Hungarian search prioritizes relaxing red-tinted admissible arcs over pure blue arcs.

We can sort the blue-tinted edges of  $\Gamma$  by the order they were relaxed into  $S$  during the Hungarian search forming  $T_i$ . Let  $(v, w) \in \Gamma$  be the last pure blue edge relaxed, of all the blue-tinted edges in  $\Gamma$  — after  $(v, w)$  is relaxed, the remaining unrelaxed, blue-tinted edges of  $\Gamma$  are purple.

Let us pause the Hungarian search the moment before  $(v, w)$  is relaxed. At this point,  $v \in S$  and  $w \notin S$ , and the Hungarian search must have finished relaxing all frontier support arcs. By our choice of  $(v, w)$ ,  $\Gamma \setminus (v, w)$  is a path of relaxed blue edges and red-tinted edges which connect  $v$  and  $w$ . Walking around  $\Gamma \setminus (v, w)$  from  $v$  to  $w$ , we see that every vertex of the cycle must be in  $S$  already:  $v \in S$ , relaxed blue edges have both endpoints in  $S$ , and any



unrelaxed red-tinted edge must have both endpoints in  $S$ , since the Hungarian search would have prioritized relaxing the red-tinted edges to grow  $S$  before relaxing  $(v, w)$  (a blue edge). It follows that  $w \in S$  already, a contradiction.

No such cycle  $\Gamma$  can exist, thus  $E(\text{supp}(f_i)) \cup E(T_i)$  is acyclic and  $E(\text{supp}(f_{i+1})) \subseteq E(\text{supp}(f_i)) \cup E(T_i)$  is acyclic. By induction,  $E(\text{supp}(f_i))$  is acyclic for all  $i$ .  $\blacktriangleleft$

Let  $E(\Sigma_a)$  **«only used once»** be the underlying edges of the support star centered at  $a$  and  $F := E(\text{supp}(f)) \setminus \bigcup_{a \in A} E(\Sigma_a)$ . Using Lemma C.6, we can show that the number of support arcs outside support stars ( $|F|$ ) is small.

► **Lemma C.7.**  $|B_\ell \setminus \bigcup_{a \in A} \Sigma_a| \leq r$ .

**Proof.**  $F$  is constructed from  $E(\text{supp}(f))$  by eliminating edges in support stars, therefore all edges in  $F$  must adjoin vertices in  $B$  of support degree at least 2. By Lemma C.6,  $E(\text{supp}(f))$  is acyclic and therefore forms a spanning forest over  $A \cup B_\ell$ , so  $F$  is also a bipartite forest. All leaves of  $F$  are therefore vertices of  $A$ .

Pick an arbitrary root for each connected component of  $F$  to establish parent-child relationships for each edge. As no vertex in  $B$  is a leaf, each vertex in  $B$  has at least one child. Charge each vertex in  $B$  to one of its children in  $F$ , which must belong to  $A$ . Each vertex in  $A$  is charged at most once. Thus, the number of  $B_\ell$  vertices outside of support stars is no more than  $r$ .  $\blacktriangleleft$

► **Lemma 5.2.** *Suppose we have stripped the graph of dead vertices. The number of relaxation steps in a Hungarian search outside of support stars is  $O(r)$ .*

**Proof.** If there are no dead vertices, then each non-support star relaxation step adds either (i) an active deficit vertex, (ii) a non-deficit vertex  $a \in A_\ell$ , or (iii) a non-deficit vertex  $b \in B_\ell$  of support degree at least 2. There is a single relaxation of type (i), as it terminates the search. The number of vertices of type (ii) is  $r$ , and the number of vertices of type (iii) is at most  $r$  by Lemma C.7. The lemma follows.  $\blacktriangleleft$

► **Lemma 5.3.** *Hungarian search takes  $O(r\sqrt{n} \text{polylog } n)$  time.*

**Proof.** The number of relaxation steps outside of support stars is  $O(r)$  by Lemma 5.2. The time per relaxation outside of support stars is  $O(\sqrt{n} \text{polylog } n)$ . The time spent processing relaxations within a support star is  $O(\sqrt{n} \text{polylog } n)$ , and at most  $r$  are relaxed during the search. The total time is therefore  $O(r\sqrt{n} \text{polylog } n)$ .  $\blacktriangleleft$

Initially, we label stars big or small according to the  $\sqrt{n}$  threshold. A star that is currently big is turned into a small star once  $|\Sigma_a| \leq \sqrt{n}/2$ . A star that is currently small is turned into a big star once  $|\Sigma_a| \geq 2\sqrt{n}$ . This way, the time spent rebuilding/updating the respective data structures can be amortized to the insertions/deletions that preceded the switch, plus some  $O(r)$  extra work if the the update is small-to-big.

A star  $\Sigma_a$  that is switching from big-to-small has size  $|\Sigma_a| \leq \sqrt{n}/2$ . When switching, we delete  $\mathcal{D}_{\text{big}}(a)$  and insert  $\Sigma_a$  into  $\mathcal{D}_{\text{small}}$ . Thus, the time spent for big-to-small update is  $O(\sqrt{n} \text{polylog } n)$ , and there were at least  $\sqrt{n}/2$  points removed from  $\Sigma_a$  since it was last big.

A star  $\Sigma_a$  that is switching from small-to-big has size  $|\Sigma_a| = \sqrt{n} + x \geq 2\sqrt{n}$ , for some integer  $x \geq \sqrt{n}$ . Rearranging, we have  $|\Sigma_a| \leq 2x$ . When switching, we delete all  $|\Sigma_a|$  points from  $\mathcal{D}_{\text{small}}$  and construct a new  $\mathcal{D}_{\text{big}}(a)$ . Constructing  $\mathcal{D}_{\text{big}}(a)$  requires inserting  $O(r)$  points of  $A$  (into  $P$ ) and the  $|\Sigma_a|$  points of the star (into  $Q$ ). Thus, the time spent for a small-to-big update is  $O((r+x) \text{polylog } n)$ , and there were at least  $x \geq \sqrt{n}$  points added to  $\Sigma_a$  since it was last small.

## 1142      ——— References for the Appendix ———

- 1144      **AFP<sup>+</sup>17** Pankaj K. Agarwal, Kyle Fox, Debmalya Panigrahi, Kasturi R. Varadarajan, and Allen  
 1145      Xiao. Faster algorithms for the geometric transportation problem. In *33rd International  
 1146      Symposium on Computational Geometry, SoCG 2017, July 4-7, 2017, Brisbane, Australia*,  
 pages 7:1–7:16, 2017.
- 1148      **Eri10** Jeff Erickson. Maximum flows and parametric shortest paths in planar graphs. In *Proceed-  
 1149      ings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA  
 2010, Austin, Texas, USA, January 17-19, 2010*, pages 794–804, 2010.
- 1151      **GHKT17** Andrew V. Goldberg, Sagi Hed, Haim Kaplan, and Robert E. Tarjan. Minimum-cost flows  
 in unit-capacity networks. *Theory Comput. Syst.*, 61(4):987–1010, 2017.
- 1153      **GT90** Andrew V. Goldberg and Robert E. Tarjan. Finding minimum-cost circulations by success-  
 1154      ive approximation. *Math. Oper. Res.*, 15(3):430–466, 1990.
- 1155      **Or193** James B. Orlin. A faster strongly polynomial minimum cost flow algorithm. *Operations  
 1156      Research*, 41(2):338–350, 1993.
- 1157      **SA12** R. Sharathkumar and Pankaj K. Agarwal. Algorithms for the transportation problem in  
 1158      geometric settings. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium  
 1159      on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012*, pages 306–317,  
 2012.
- 1160      **Vai89** Pravin M. Vaidya. Geometry helps in matching. *SIAM J. Comput.*, 18(6):1201–1225, 1989.