Better bounds for matchings in the streaming model

Michael Kapralov
EPFL

March 17, 2021

Abstract

In this paper we present improved bounds for approximating maximum matchings in bipartite graphs in the streaming model. First, we consider the question of how well maximum matching can be approximated in a single pass over the input when $\tilde{O}(n)$ space is allowed, where n is the number of vertices in the input graph. Two natural variants of this problem have been considered in the literature: (1) the edge arrival setting, where edges arrive in the stream and (2) the vertex arrival setting, where vertices on one side of the graph arrive in the stream together with all their incident edges. The latter setting has also been studied extensively in the context of *online algorithms*, where each arriving vertex has to either be matched irrevocably or discarded upon arrival. In the online setting, the celebrated algorithm of Karp-Vazirani-Vazirani achieves a $1 - 1/e$ approximation by crucially using randomization (and using $\tilde{O}(n)$ space). Despite the fact that the streaming model is less restrictive in that the algorithm is not constrained to match vertices irrevocably upon arrival, the best known approximation in the streaming model with vertex arrivals and $\tilde{O}(n)$ space is the same factor of $1 - 1/e$.

We show that no (possibly randomized) single pass streaming algorithm constrained to use $\tilde{O}(n)$ space can achieve a better than $1 - 1/e$ approximation to maximum matching, even in the vertex arrival setting. This leads to the striking conclusion that no single pass streaming algorithm can get any advantage over online algorithms unless it uses significantly more than $\tilde{O}(n)$ space. Additionally, our bound yields the best known impossibility result for approximating matchings in the *edge arrival* model (improving upon the bound of $2/3$ proved by Goel at al[SODA'12]).

Second, we consider the problem of approximating matchings in multiple passes in the vertex arrival setting. We show that a simple fractional load balancing approach achieves approximation ratio $1 - e^{-k}k^{k-1}/(k-1)! = 1 - \frac{1}{\sqrt{2\pi k}} + o(1/k)$ in k passes using linear space. Thus, our algorithm achieves the best possible $1 - 1/e$ approximation in a single pass and improves upon the $1 - O(\sqrt{\log\log k/k})$ approximation in k passes due to Ahn and Guha[ICALP'11]. Additionally, our approach yields an efficient solution to the Gap-Existence problem considered by Charles et al[EC'10].

1 Introduction

The need to process modern massive data sets necessitates rethinking classical solutions to many combinatorial optimization problems from the point of view of space usage and type of access to the data that algorithms assume. Applications in domains such as processing web-scale graphs, network monitoring or data mining among many others prohibit solutions that load the whole input into memory and assume random access to it. The streaming model of computation has emerged as a more realistic model for processing modern data sets. In this model the input is given to the algorithm as a stream, possibly with multiple passes allowed. The goal is to design algorithms that require small space and ideally one or a small constant number of passes over the data stream to compute a (often approximate) solution. For many problems with applications in network monitoring, it has been shown that space polylogarithmic in the size of the input is often sufficient to compute very good approximate solutions. On the other hand, even basic graph algorithms have been shown to require $\Omega(n)$ space in the streaming model[FKM$^+$05], where n is the number of vertices. A common relaxation is to allow $O(n \cdot \text{polylog}(n))$ space, a setting often referred to as the *semi-streaming* model.

1.1 Matchings in the streaming model

The problem of approximating maximum matchings in bipartite graphs has received significant attention recently, and very efficient small-space solutions are known when multiple passes are allowed[FKM$^+$04, McG05, EKS09, AG11, KMM12]. The best known algorithm due to Ahn and Guha [AG11] achieves a $1 - O(\sqrt{\log\log k/k})$ in k passes for the weighted as well as the unweighted version of the problem using $\tilde{O}(kn)$ space.

All algorithms mentioned above require at least two passes to achieve a nontrivial approximation. The problem of approximating matchings in a single pass has recently received significant attention[GKK12, KMM12]. Two natural variants of this problem have been considered in the literature: (1) the edge arrival setting, where edges arrive in the stream and (2) the vertex arrival setting, when vertices on one side of the graph arrive in the stream together with all their incident edges. The latter setting has also been studied extensively in the context of *online algorithms*, where each arriving vertex has to either be matched irrevocably or discarded upon arrival.

In a single pass, the best known approximation in the edge arrival setting is still $1/2$, achieved by simply keeping a maximal matching (this was improved to $1/2 + \epsilon$ for a constant $\epsilon > 0$ under the additional assumption of random edge arrivals [KMM12]). It was shown in [GKK12] that no $\tilde{O}(n)$ space algorithm can achieve a better than $2/3$ approximation in this setting.

In the vertex arrival setting, the best known algorithms achieve an approximation of $1 - 1/e$. The assumption of vertex arrivals allows one to leverage results from online algorithms [KVV90, MY11, KMT11]. In the online model vertices on one side of the graph are known, and vertices on the other side arrive in an adversarial order. The algorithm has to either match a vertex irrevocably or discard upon arrival. The celebrated algorithm of Karp-Vazirani-Vazirani achieves a $1 - 1/e$ approximation for the online problem by crucially using randomization (additionally, this algorithm only uses $\tilde{O}(n)$ space). A *deterministic* single pass $\tilde{O}(n)$ space $1 - 1/e$ approximation in the vertex arrival setting was given in [GKK12] (such a deterministic solution is provably impossible in the online setting). In [GKK12], the authors also showed by analyzing a natural one-round communication problem that no single-pass streaming algorithm that uses $\tilde{O}(n)$ space can obtain a better than $3/4$ approximation in the vertex arrival setting. They also provided a protocol for this communication problem that matches the $3/4$ approximation ratio, suggesting that new techniques would be needed to prove a stronger impossibility result.

Recent work. The lower bound presented in this paper has recently been improved to $\frac{1}{1+\ln 2} \approx 0.591$ by [Kap21] for the more general edge arrival model, following exciting developments in online match-

ing [WW15, ELSW13, GKM$^+$19]. A $2/3$-approximation to maximum matching in a single pass over a randomly ordered stream of edges in $n \log^{O(1)} n$ space has recently been given by [Ber20] (reducing the space complexity of the approach of [ABB$^+$19] from $\widetilde{O}(n^{3/2})$ to $n \log^{O(1)} n$), and very recently improved to $2/3 + \Omega(1)$ by [AB21]. A $1 - O(1/\sqrt{k})$ approximation in k passes using $n \log^{O(1)} n$ space was given by [ALT21].

1.2 Our results

In this paper, we improve upon the best known bounds for both the single pass and multi-pass settings. In the single pass setting, we prove an optimal impossibility result for vertex arrivals, which also yields the best known impossibility result in the edge arrival model. For the multipass setting, we give a simple algorithm that improves upon the approximation obtained by Ahn and Guha in the vertex arrival setting, as well as yields an efficient solution to the Gap-Existence problem considered by Charles et al[CCD$^+$10].

Lower bounds. Our main result is an optimal bound on the best approximation ratio that a single-pass $\tilde{O}(n)$ space streaming algorithm can achieve in the vertex arrival setting:

Theorem 1 *No (possibly randomized) one-pass streaming algorithm can obtain a $(1-1/e+c)$-approximation to the maximum matching with probability at least $3/4$ for any constant $c > 0$, unless it uses at least $n^{1+\Omega_c(1/\log\log n)}$ space, even in the vertex arrival model.*

Remark 2 *In fact, we prove a more refined statement: for every integer $k \geq 2$ if the edge set is partitioned among k players communicating in the number-in-hand model (with the i-th player sending a single message to the $(i+1)$-th after receiving a message from the $(i-1)$-th player) no algorithm can achieve a $1 - (1 - 1/k)^k + \Omega(1)$ approximation to maximum matching unless it uses $n^{1+\Omega(1/\log\log n)}$ communication.*

We note that this bound is matched by the randomized KVV algorithm[KVV90] for the online problem and the deterministic $\tilde{O}(n)$ space algorithm of [GKK12]. One striking consequence of our bound is that no single-pass streaming algorithm can improve upon the more constrained *online* algorithm of KVV, which has to make irrevocable decisions, unless is uses significantly more than $\tilde{O}(n)$ space. Our bound also improves upon the best known bound of $2/3$ for small space one-pass streaming algorithms in the *edge arrival model*.

It was shown in [GKK12] via an analysis of the natural two-party communication problem that no one-pass streaming algorithm that uses $\tilde{O}(n)$ space can achieve approximation better than $2/3$ in the edge arrival setting and $3/4$ in the vertex arrival setting. Furthermore, the authors also gave a communication protocol that proves the optimality of both bounds for the communication problem, thus suggesting that a more intricate approach would be needed to prove better impossibility results. While the lower bounds from [GKK12] follow from a construction of a distribution on inputs that consists of two parts and hence yields a two-party communication problem, here we obtain an improvement by constructing hard input sequences that consist of k parts instead of two, getting a lower bound that approaches $1 - 1/e$ for large k.

Upper bounds. We show that a simple algorithm based on fractional load balancing achieves the optimal $1 - 1/e$ approximation in a single pass and $1 - \frac{1}{\sqrt{2\pi k}} + o(k^{-1/2})$ approximation in k passes, improving upon the best known algorithms for this setting:

Theorem 3 *There exists an algorithm for approximating the maximum matching M in a bipartite graph $G = (P, Q, E)$ with the P side arriving in the stream to factor $1 - e^{-k}k^{k-1}/(k-1)! = 1 - \frac{1}{\sqrt{2\pi k}} + O(k^{-3/2})$ in k passes using $O(|P| + |Q|)$ space. The algorithm can be implemented to run in nearly linear time in the number of edges in the graph per pass, with space complexity $\widetilde{O}(|P| + |Q|)$.*

The gap-existence problem. In [CCD$^+$10] the authors give an algorithm for the closely related *gap-existence* problem. In this problem the algorithm is given a bipartite graph $G = (A, I, E)$, where A is the set of advertisers with budgets $B_a, a \in A$ and I is the set of impressions. The graph is lopsided in the sense that $|I| \gg |A|$. A matching M is *complete* if $|M \cap \delta(i)| = 1$ for all $i \in I$ and $|M \cap \delta(a)| = B_a$ for all $a \in A$. The gap-existence problem consists of distinguishing between two cases:

(YES) there exists a complete matching with budgets B_a;

(NO) there does not exist a complete matching with budgets $\lfloor (1-\epsilon)B_a \rfloor$.

The approach of [CCD$^+$10] is via sampling the I side of the graph, and yields a solution that allows for non-trivial subsampling when the budgets are large. In particular, they obtain an algorithm with runtime $O\left(\frac{|A|\log|A|}{\epsilon^2} \cdot \frac{|I|}{\min_a |B_a|}\right)$, which is sublinear in the size of the graph when all budgets are large. In Section 5 we improve significantly upon their result, showing

Theorem 4 *Gap-Existence can be solved in $O(\log(\frac{1}{\epsilon}\sum_{a\in B_a} B_a)/\epsilon^2)$ passes using space $O(\sum_{a\in A} B_a/\epsilon)$. The time taken for each pass is nearly linear in the representation of the graph.*

It should also be noted that the result of [CCD$^+$10] could be viewed as a single pass algorithm, albeit with the stronger assumption that the arrival order in the stream is random.

Organization: We start by presenting a toy version of our lower bound construction in Section 2. The construction in Section 2 does not give a strong streaming lower bound, but captures most of the properties of our hard input distribution, while at the same time being quite simple to describe. In Section 3 we give the actual lower bound construction and prove Theorem 1. Our basic multipass algorithm for approximating matchings is presented in Section 4, and the algorithm for Gap-existence is given in Section 5.

2 A toy construction

In this section we show that for every integer $k \geq 2$ there exists a distribution $\mathcal{D}$ on input instances to the bipartite matching problem such that a graph G with N vertices sampled from distribution $\mathcal{D}$ has a nearly perfect matching with high probability, but any single-pass streaming algorithm that maintains a subset of edges of G in memory and outputs a matching in the subset of edges retained cannot achieve a better than $1-(1-1/k)^k+\delta$ approximation for a constant $\delta > 0$ unless it maintains $\Omega(N \log N)$ edges.

We define a family of graphs that forms the basis of our hard input instances in Section 2.1. In Section 2.2 we define a hard input distribution based on these graphs, prove Theorem 15 (our main result in this section), which provides the $1-(1-1/k)^k+\delta$ upper bound on the approximation ratio that an algorithm that stores $o(n \log n)$ edges.

2.1 Construction of the input family of graphs

We construct bipartite graphs $G = (S, T, E)$, with S and T the two sides of the bipartition.

Vertices of G: the T side of the bipartition Let $k \geq 2$ be a large constant integer. Let $m \geq 1$ a multiple of k be a sufficiently large integer. Let $T = [m]^n$, i.e. vertices in T are vectors of dimension n, with each co-ordinate taking values in $[m] = \{1, 2, \ldots, m\}$. This way we have $N := |T| = m^n$, so $n = \Omega(\log N)$ for every constant m. The vertices on the S side of the bipartition will also be associated with points on the hypercube $[m]^n$, as defined below.

Vertices of G: the S side of the bipartition To define the vertices in the partition $S = S_0 \cup S_1 \cup \ldots \cup S_k$, we first partition the set of coordinates $[n]$ into k equal size blocks $[n] = B_1 \cup \ldots \cup B_k$. Graphs $G = G(j_1, \ldots, j_k)$ will be parameterized by a sequence $(j_1, \ldots, j_k) \in B_1 \times \ldots \times B_k$ of coordinates. Also for each point $x \in [m]^n$ let Z_x be an independent Bernoulli 0/1 random variable with expectation $1/k$ – we will later choose some fixing of these random variables for the final construction. Then for every $i = 0, \ldots, k$ we let

$$\begin{aligned} T_i &= \{y \in [m]^n : y_{j_r} \in (m/k, m] \text{ for all } r = 1, \ldots, i\} \\ S_i &= \{x \in T_i : Z_x = 1\}. \end{aligned} \tag{1}$$

Note that $T_0 = T$, and for every $i = 0, \ldots, k-1$ the set S_i is a subsampling of T_i at rate $1/k$. We also let, for every $i = 0, \ldots, k-1$ and $j \in B_{i+1}$

$$\begin{aligned} T_i^j &= \{y \in T_i : y_j \in (m/k, m]\} \\ S_i^j &= \{x \in S_i : Z_x = 1 \text{ and } x_j \in (m/k, m]\}. \end{aligned}$$

We also define for each $i = 0, \ldots, k-1$

$$S_i^* = \{x \in S_i : x_{j_r} \in (m/k, m] \text{ for all } r = i+1, \ldots, k\}. \tag{2}$$

We will use

Theorem 5 (Chernoff bound) *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables, let $\mu := \mathbb{E}[\sum_{i=1}^n X_i]$. Then for every $\delta \in (0,1)$ one has $\mathbb{P}[|\sum_{i=1}^n X_i - \mu| > \delta\mu] \leq 2e^{-\delta^2\mu/3}$.*

We first note that

Lemma 6 *For any $k \geq 2$ the following conditions hold.* **(1)** *For every choice of $j_1, \ldots, j_k$ and every $i = 0, \ldots, k$ one has $|T_i| = (1 - \frac{1}{k})^i |T|$. For every $\eta \in (0, 1/2)$ there exists an event $\mathcal{E}_{set-sizes}$ that occurs with probability at least $1 - k(\log N)^k e^{-\Omega(\eta^2 N/k)}$ over the random variables $Z_x, x \in [m]^n$ such that conditioned on $\mathcal{E}_{set-sizes}$ one has for every choice of $(j_1, \ldots, j_k) \in B_1 \times \ldots \times B_k$ simultaneously for every $i = 0, \ldots, k-1$* **(2)** *$|S_i| = (1 \pm \eta)|T_i|/k$,* **(3)** *$|S_i^j| = (1 \pm O(\eta))(1 - 1/k)|S_i|$, and* **(4)** *$|S_i^*| = (1 \pm \eta)|T_k|/k$ (note that this quantity does not depend on i).*

Proof: **(1)** follows directly by definition of T_i. For **(2)** we first note that by an application of Chernoff bounds for a fixed collection $j_1, \ldots, j_k$ one has $|S_i| = (1 \pm \eta)|T_i|/k$ with probability at least $1 - e^{-\Omega(\eta^2 N/k)}$, where we used the fact that $(1-1/k)^i \geq (1-1/k)^k \geq (1-1/2)^2$ for every $i = 0, \ldots, k$, since $k \geq 2$ by assumption of the lemma. A union bound over at most $(\log N)^k$ choices for $j_1, \ldots, j_k$ and k choices for i gives the result of the lemma. The third and fourth bound follow analogously. ■

We need the following simple lemma:

Lemma 7 *For every $i = 0, \ldots, k-1$, every $(j_1, \ldots, j_i) \in B_1 \times \ldots \times B_i$ the following conditions hold. For every $j \in B_{i+1}$, every $z \in T_i^j$ let $\deg_j(z)$ denote the number of $j' \in B_{i+1} \setminus \{j\}$ such that $z \in T_i^{j'}$. Let $\overline{\deg}_j(z)$ denote the number of $j' \in B_{i+1} \setminus \{j\}$ such that $z \in T_i \setminus T_i^j$. Then for every $\eta \in (0, 1/2)$ one has $\deg_j(z) \in (1 \pm \eta)(1 - 1/k)(|B_{i+1}| - 1)$ and $\overline{\deg}_j(z) \in (1 \pm \eta)(|B_{i+1}| - 1)/k$ for all but a $N^{-\Omega(\eta^2/m^2)}$ fraction of $z \in T$. The same bounds hold for $z \in S_i^j$.*

Proof: Recall that $T_i = \{y \in [m]^n : y_{j_r} \in (m/k, m] \text{ for all } r = 1, \ldots, i\}$. We thus have

$$T_i^j = \{y \in [m]^n : y_{j_r} \in (m/k, m] \text{ for all } r = 1, \ldots, i \text{ and } y_j \in (m/k, m]\}$$

and

$$T_i^{j'} = \left\{ y \in [m]^n : y_{j_r} \in (m/k, m] \text{ for all } r = 1, \ldots, i \text{ and } y_{j'} \in (m/k, m] \right\}.$$

Since $j_r \in B_r$ for every $r = 1, \ldots, i$, and $j, j' \in B_{i+1}$, and $B_1, \ldots, B_k$ are disjoint, we have that coordinate $y_{j'}$ is unconstrained in T_i^j, a uniformly random $z \in T_i^j$ satisfies $z_{j'} \in (m/k, m]$ with probability exactly $1 - 1/k$. Furthermore, these events are independent for different collections of coordinates in $B_{i+1} \setminus \{j\}$. Select $z \in T_i^j$ uniformly at random. For $j' \in B_{i+1} \setminus \{j\}$ let $F_{j'} = 1$ if $z \in T_i^{j'}$ and $F_{j'} = 0$ otherwise (note that $\mathbb{E}[F_{j'}] = 1 - 1/k$ for every $j' \in B_{i+1} \setminus \{j\}$). We now have by the Chernoff bound (Theorem 5) that for every $\eta \in (0, 1/2)$

$$\mathbb{P}_{z \sim UNIF(T_i^j)} \left[\sum_{j' \in B_{i+1} \setminus \{j\}} F_{j'} \notin (1 \pm \eta)(1 - 1/k)(|B_{i+1}| - 1) \right] \leq 2e^{-\Omega(\eta^2 |B_{i+1}|)} = N^{-\Omega(\eta^2/m^2)},$$

where we used the fact that $|B_{i+1}| = n/m = (\log_m N)/m$ in the last transition. This proves the first claim. The proof of the second and third claim is analogous. ■

Edges of G. For each $i = 0, \ldots, k-1$ edges of the subgraph $G = (P_i, Q, E_i)$ will be associated with coordinates in B_{i+1}, as we now describe. Specifically, each coordinate $j \in B_{i+1}$ will correspond to a set of edges in G that form a rather large near-matching (of size $\Omega(N/k)$, as described below).

For each $i = 0, \ldots, k-1$ the edge set $E_i \subseteq S_i \times T_i$ are defined as follows. For each coordinate $j \in B_i$ for each $x \in [m]^n$ we let

$$\text{line}_j(x) = \{x' \in [m]^n : (x' - x)_s = 0 \text{ for all } s \neq j\}$$

denote the line through x in coordinate direction j. Note that $|\text{line}_j(x)| = m$ for all x. Furthermore, we have

Lemma 8 *For every $\eta \in (0, 1/2)$, if $C > 0$ is a sufficiently large constant, then for $m \geq C\eta^{-2} k \log \eta^{-1}$ a multiple of k, for every $i = 0, \ldots, k-1$, every $(j_1, \ldots, j_i) \in B_1 \times \ldots \times B_i$ for each $y \in T_i$ one has for each $j \in B_{i+1}$*

(1) $|\text{line}_j(y)| = m$ and $\text{line}_j(y) \subseteq T_i$;

(2) $|\text{line}_j(y) \setminus T_i^j| = m/k$;

(3) there exists an event $\mathcal{E}_{large-lines}(j_1, \ldots, j_i, j)$ that occurs with probability at least $1 - e^{-\Omega(\eta^2 N/k)}$ such that conditioned on $\mathcal{E}_{large-lines}(j_1, \ldots, j_i, j)$ the number of $y \in T_i$ such that $|\text{line}_j(y) \cap S_i^j| \notin (1 \pm \eta)|\text{line}_j(y)|(1 - 1/k)/k$ is upper bounded by $\eta^2 |T_i|$.

In particular, there exists an event $\mathcal{E}_{large-lines}$ that occurs with probability at least $1 - k(\log N)^k e^{-\Omega(\eta N/k)}$ such that for every $i = 0, \ldots, k-1$, every collection $j_1, \ldots, j_i$, every $j \in B_{i+1}$ one has that the number of $y \in T_i$ such that $|\text{line}_j(y) \cap S_i| \notin (1 \pm \eta)|\text{line}_j(x)|/k$ is upper bounded by $2\eta^2 |T_i|$.

Proof: The first claim follows since, due to the assumption that $y \in T_i$ we have

$$\begin{aligned} \text{line}_j(y) &= \left\{ y' \in [m]^n : (y' - y)_s = 0 \text{ for all } s \neq j \right\} \\ &= \left\{ y' \in [m]^n : (y' - y)_s = 0 \text{ for all } s \neq j, y'_{j_r} \in (m/k, m] \text{ for all } r = 1, \ldots, i \right\} \\ &\subseteq T_i \end{aligned}$$

since $j \neq j_1, \ldots, j_i$ due to the assumption that $j \in B_{i+1}$.

The second claim follows similarly. For the third claim note that

$$\mathbb{E}_Z\left[|\text{line}_j(y)\cap S_i^j|\right] = \sum_{y'\in\text{line}_j(y)\cap T_i^j}\mathbb{P}_Z[y\in S_i^j]$$
$$= |\text{line}_j(x)|(1-1/k)/k,$$

where we used the fact that $|\text{line}_j(y)\cap T_i^j| = m/k$ for every $y\in T_i$ by **(2)** and $|\text{line}_j(y)| = m$ by **(1)**. Since $m/k \geq C\eta^{-2}\log\eta^{-1}$ for a constant $C>0$ by assumption of the lemma, the claim follows by the Chernoff bound (Theorem 5). The final claim follows by a union bound over all choices of $i, j_1, \ldots, j_i, j$. ■

We now condition on the event $\mathcal{E}_{large-lines}$ from Lemma 8, so that that $|\text{line}_j(x)\cap S_i^j| \notin (1\pm\eta)|\text{line}_j(x)|/k$ for all $i=0,\ldots,k-1$, $j\in B_{i-1}$ and all but $2\eta^2|T_i|$ choices of $x\in T_i$.

Defining the edges induced by $T_i\cup S_i$. We now define the edges of $G = G(j_1,\ldots,j_k)$ induced by $T_i\cup S_i$ (note that these edges are a function of the prefix $(j_1,\ldots,j_i)$ only). The edge set is a union of a large number of induced subgraphs of constant size. We will need

Definition 9 (Typical line) *For every $i=0,\ldots,k-1$, every $(j_1,\ldots,j_i)\in B_1\times\ldots\times B_i$, $j\in B_{i+1}$, for $z\in T_i$ we say that line$_j(z)$ is typical if $|\text{line}_j(z)\cap S_i^j|\in(1\pm\eta)|\text{line}_j(z)|(1-1/k)/k = (1\pm\eta)(1-1/k)m/k$ and atypical otherwise.*

For every $y\in T_i$, if $\text{line}_j(y)$ is typical, let $\widetilde{\text{line}}_j(y)$ be an arbitrary subset of $\text{line}_j(y)\cap S_i^j$ of size $(1-\eta)|\text{line}_j(y)|(1-1/k)/k = (1-\eta)(1-1/k)\cdot m/k$, and let $\widetilde{\text{line}}_j(y) := \emptyset$ otherwise. We now define the edge set of E_i. For every $j\in B_{i+1}$, every $y\in T_i$ include a complete bipartite graph between $\widetilde{\text{line}}_j(y)$ and $\text{line}_j(y)\cap(T_i\setminus T_i^j)$, i.e.

$$E_i = \bigcup_{j\in B_{i+1}} E_i^j, \text{ where } E_i^j = \bigcup_{y\in T_i}\widetilde{\text{line}}_j(y)\times(\text{line}_j(y)\cap(T_i\setminus T_i^j)). \tag{3}$$

Note that for every $a\in\text{line}_j(y)\cap(T_i\setminus T_i^j)$ and $b\in\widetilde{\text{line}}_j(y)$ we have $(a-b)_q = 0$ for all $q\neq j$, $a_j\in[1,m/k]$ and $b_j\in(m/k,m]$. We now prove that for every j there exists a matching of (most of) S_i to $T_i\setminus T_i^j$.

First note that it follows immediately that there exists a matching of at least a $(1-1/k-O(\eta+\eta^2k))$ fraction of S_i to $T_i\setminus T_i^j$. Indeed, for every $y\in T_i\setminus T_i^j$ such that $\text{line}_j(y)$ is typical as per Definition 9 one can match $\widetilde{\text{line}}_j(y)$, which constitutes a $(1-\eta)(1-1/k)$ fraction of $\text{line}_j(y)$, to $\text{line}_j(y)\cap(T_i\setminus T_i^j)$ through the edges of the complete bipartite graph $\widetilde{\text{line}}_j(y)\times(\text{line}_j(y)\cap(T_i\setminus T_i^j))$. At the same time the number of y's that belong to atypical lines is at most $2\eta^2|T_i| = O(\eta^2k)|S_i|$ by conditioning on $\mathcal{E}_{large-lines}$ and the high probability event $\mathcal{E}_{set-sizes}$ from Lemma 6. While this would have sufficed for proving a $1-1/e$ lower bound, we would like to get a lower bound of $1-(1-1/k)^k$ for every $k\geq 2$. For that we need the slightly harder

Lemma 10 *For every $\eta\in(0,1/2)$, if $C>0$ is a sufficiently large constant, then for $m\geq C\eta^{-2}k\log\eta^{-1}$ a multiple of k, conditioned on $\mathcal{E}_{large-lines}$ (defined in Lemma 8) and $\mathcal{E}_{set-sizes}$ (defined in Lemma 6) for every $i=0,\ldots,k-1$, every $(j_1,\ldots,j_i)\in B_1\times\ldots\times B_i$ for each $j\in B_{i+1}$ there exists a matching of at least $(1-O(\eta+\eta^2k))|S_i| - N^{-\Omega(\eta^2/m^2)}$ nodes in S_i to $T_i\setminus T_i^j$ for sufficiently large N.*

Proof: Let $C>0$ be sufficiently large as prescribed by Lemma 8. We prove the existence of the required matching by exhibiting a fractional matching of appropriate size, which implies the result by the integrality of the bipartite matching polytope. The construction proceeds over three steps.

Step 1 For every $x \in S_i$ such that $\text{line}_j(x)$ is typical put fractional mass k/m on every edge in $\widetilde{\text{line}_j(x)} \times (\text{line}_j(x) \cap (T_i \setminus T_i^j))$. Since $|\text{line}_j(x) \cap (T_i \setminus T_i^j)| = m/k$ by Lemma 8, **(2)**, this places a unit of mass on the neighborhood of every vertex in $\widetilde{\text{line}_j(x)}$. Since $|\widetilde{\text{line}_j(x)}| = (1-\eta)(1-1/k) \cdot m/k$ by definition, this places fractional mass $(1-\eta)(1-1/k)$ on every $y \in \text{line}_j(x) \cap (T_i \setminus T_i^j)$, leaving at least $1/k$ capacity on each such y. We assign more fractional mass to use the remaining $1/k$ mass up to an $O(\eta)$ term in step 2.
Step 2 For every $x \in S_i \setminus S_i^j$ put fractional mass

$$\epsilon := \frac{1}{(m/k) \cdot (1+\eta)(1-1/k)(|B_j|-1)} \tag{4}$$

on every edge connecting x to $y \in T_i$. Note that these edges correspond to coordinates $j' \in B_{i+1} \setminus \{j\}$. In particular, if (x,y) is an edge corresponding to coordinate j', then we have $y_q = x_q$ for all $q \neq j'$, and in particular it must be that $y \in T_i \setminus T_i^j$.
Step 3 Let $\deg_j(x)$ denote the number of $j' \in B_{i+1} \setminus \{j\}$ such that $x \in T_i^{j'}$, and let $\overline{\deg_j}(y)$ denote the number of $j' \in B_{i+1} \setminus \{j\}$ such that $y \in T_i \setminus T_i^j$. We now remove all fractional mass assigned to vertices $x \in S_i$ with $\overline{\deg_j}(x) \notin (1 \pm \eta)\frac{1}{k}(|B_{i+1}|-1)$ and vertices $y \in T_i$ with $\deg_j(z) \notin (1 \pm \eta)(1-1/k)(|B_{i+1}|-1)$. We refer to such nodes as *atypical*.

We now prove upper and lower bounds on the fractional mass assigned by this rule to every $x \in S_i^j, y \in T_i \setminus T_i^j$. This establishes feasibility of the fractional solution and lower bounds its value respectively.

Upper bounding load (feasibility). For every $j' \in B_{i+1} \setminus \{j\}$ every vertex x is either connected to exactly $|\text{line}_j(x) \cap (T_i \setminus T_i^j)| = m/k$ nodes in $T_i \setminus T_i^j$ with edges in $E_i^{j'}$ or zero nodes (when x belongs to an atypical line in direction j'). In the former case coordinate j' contributes exactly $\epsilon \cdot (m/k)$ fractional mass (where ϵ is defined in (4)), and in the latter it contributes 0. We now get that the total mass contributed to x by directions $j' \neq j$ is no larger than $\deg_j(x) \cdot (m/k) \cdot \epsilon = \deg_j(x) \cdot (m/k) \cdot \frac{1}{(1+\eta)(m/k) \cdot (1-1/k)(|B_j|-1)}$. By Lemma 7 for all but $N^{1-\Omega(\eta^2/m^2)}$ of $x \in S_i$ one has

$$\deg_j(z) \in (1 \pm \eta)(1-1/k)(|B_{i+1}|-1). \tag{5}$$

We call such x typical. We thus get that the total mass assigned to edges incident on typical $x \in S_i \setminus S_i^j$ is upper bounded by $(1+\eta)(1-1/k)(|B_{i+1}|-1) \cdot (m/k) \cdot \frac{1}{(m/k)(1+\eta) \cdot (1-1/k)(|B_j|-1)} \leq 1$, and the fractional assignment is feasible for all but $N^{1-\Omega(\eta^2/m)}$ nodes (i.e. for all typical nodes as per definition above).

Similarly, get by Lemma 7 for all but $N^{1-\Omega(\eta^2/m^2)}$ of $y \in T_i \setminus T_i^j$ one has

$$\overline{\deg_j}(y) \in (1 \pm \eta)\frac{1}{k}(|B_{i+1}|-1). \tag{6}$$

Now note that $|\widetilde{\text{line}_j}(x)| = (1-\eta)m(1-1/k)/k$ for every x such that the corresponding line is typical. The degree in E_i^j of a vertex $y \in T_i \setminus T_i^j$ such that $\text{line}_{j'}(y)$ is thus exactly $(1-\eta)m(1-1/k)/k$ if the corresponding line is typical, and is zero otherwise. The amount of mass assigned to y is thus $\overline{\deg_j}(y) \cdot ((1-\eta)m(1-1/k)/k) \cdot \frac{1}{(1+\eta)(m/k) \cdot (1-1/k)(|B_j|-1)} \leq (1-\eta)/k \leq 1/k$. Thus, together with the amount of mass assigned in **Step 1** to vertices $y \in T_i \setminus T_i^j$, our assignment is feasible for all but $N^{1-\Omega(\eta^2/m)}$ nodes (i.e. for all typical nodes as per definition above).

Lower bounding fractional matching size. In Step 1 we assigned $(1-\eta)(1-1/k)$ to every node in $y \in T_i \setminus T_i^j$ that belongs to a typical line in direction j. The number of such nodes is at least $(1 - O(\eta^2 k))|S_i|$ by Lemma 8, **(3)** together with Lemma 6, since we condition on $\mathcal{E}_{large-lines}$ and $\mathcal{E}_{set-sizes}$. In Step 2 we

assigned $\epsilon := \frac{1}{(m/k)\cdot(1+\eta)(1-1/k)(|B_j|-1)}$ mass to every edge from $x \in S_i \setminus S_i^j$ to $y \in T_i \setminus T_i^j$ along some direction $j' \in B_{i+1} \setminus \{j\}$ if the corresponding line is typical. Thus, for every j' we assigned $\epsilon \cdot (m/k)$ mass to every x that belonged to a typical line in direction j' (all but $O(\eta^2 k)|S_i|$ such x for every direction j' by conditioning on $\mathcal{E}_{large-lines}$). Altogether $x \in S_i \setminus S_i^j$ thus contributed at least

$$\begin{aligned}
&\sum_{j' \in B_{i+1}\setminus\{j\}} \sum_{\substack{x \in S_i \setminus S_i^j : x \text{ typical and} \\ \text{line}_{j'}(x) \text{ typical}}} \epsilon \cdot (m/k) \\
&\geq \sum_{j' \in B_{i+1}\setminus\{j\}} \left(-\epsilon \cdot (m/k) \cdot \eta^2 |T_i| + \sum_{x \in S_i \setminus S_i^j : x \text{ typical}} \epsilon \cdot (m/k) \right) \\
&= -\eta^2 \epsilon (m/k) \cdot |B_{i+1}| \cdot |T_i| + \sum_{x \in S_i \setminus S_i^j : x \text{ typical}} \epsilon \cdot (m/k) \cdot \deg_j(x) \\
&= -O(\eta^2)|T_i| + \sum_{x \in S_i \setminus S_i^j : x \text{ typical}} \epsilon \cdot (m/k) \cdot \deg_j(x),
\end{aligned}$$

where we used the fact that, conditioned on $\mathcal{E}_{large-lines}$, by Lemma 8, **(3)** for every i and every $j \in B_{i+1}$ all but $\eta^2|T_i|$ belong to typical lines in direction j, as well as the definition of ϵ in (4). We now lower bound the second term:

$$\begin{aligned}
\sum_{x \in S_i \setminus S_i^j : x \text{ typical}} \epsilon \cdot (m/k) \cdot \deg_j(x) &\geq \sum_{x \in S_i \setminus S_i^j : x \text{ typical}} \epsilon \cdot (m/k)(1-\eta)(1-1/k)(|B_{i+1}|-1) \\
&\geq \sum_{x \in S_i \setminus S_i^j : x \text{ typical}} (1 - O(\eta)) \\
&\geq (1 - O(\eta))|S_i \setminus S_i^j| - N^{-\Omega(\eta^2/m^2)},
\end{aligned}$$

where the first transition is by definition of typical x, and the second is by Lemma 7. Putting the bounds above together shows that we constructed a fractional matching of size at least $(1 - O(\eta + \eta^2 k))|S_i| - N^{-\Omega(\eta^2/m^2)}$, as required. ■

2.2 Hard input distribution and its analysis

Hard input distribution. First select values of random variables $\{Z_x\}_{x \in [m]^n}$ so that $\mathcal{E}_{set-sizes}$ and $\mathcal{E}_{large-lines}$ occur (we will verify that this is feasible later in the proof of Theorem 1, where we set parameters). The input graph is generated as follows. First for every $i = 0, \ldots, k-1$ let j_i be uniformly random in B_i. Then for each $i = 0, \ldots, k-1$ the edges of the graph induced by $S_i \cup T_i$, namely E_i (defined in (3)) arrive in the stream in an arbitrary order. Finally, a perfect matching of T_k to a fresh set S_k of vertices on the S side arrives. We denote this distribution over input graphs by $\mathcal{D}$. In this section we are assuming a stylized model, where after every stage the algorithm must select $s = o(N \log N)$ edges to keep in memory, and at the end of the stream must output a matching in the subgraph that it maintained. We show in Theorem 1 that no such algorithm can achieve a better than $1 - 1/e$ approximation to maximum matching. More specifically, we show that no algorithm can achieve a significantly better than factor $1 - (1 - 1/k)^k$ approximation on a k-stage input instance for every constant $k \geq 2$.

Intuition for the construction and lower bound. We will show in that in order to have performance better than $1-(1-1/k)^k+\delta$ on our instance the algorithm needs to store at least $\Omega(\delta N/k)$ edges from at least one of the sets $E_i^{j_{i+1}}$ (see (3)), for some $i=0,\ldots,k-1$. However, since at each step j_{i+1} is uniformly random in B_{i+1} this is impossible if the algorithm can only store $s=o(N\log N)$ edges (i.e. any sublinear fraction of the total number of edges in the graph).

The analysis relies on the several auxiliary lemmas. First, we show that the input graph contains a large matching:

Lemma 11 *For every $\eta\in(0,1/2)$, if $C>0$ is a sufficiently large constant, then for $m\geq C\eta^{-2}k\log\eta^{-1}$ a multiple of k, conditioned on $\mathcal{E}_{large-lines}$ (defined in Lemma 8) and $\mathcal{E}_{set-sizes}$ (defined in Lemma 6), every $(j_1,\ldots,j_k)\in B_1\times\ldots\times B_k$ the graph $G=G(j_1,\ldots,j_k)$ contains a matching of size at least $(1-O(\eta+\eta^2k))|S|$ if N is sufficiently large.*

Proof: Let $C>0$ be sufficiently large as dictated by Lemma 10. Now by Lemma 10 for every $i=0,\ldots,k-1$ match at least $(1-O(\eta+\eta^2k))|S_i|-N^{-\Omega(\eta^2/m^2)}|T|$ of S_i to $T_i\setminus T_i^j$. Then match S_k to T_k. For every fixed k,η,m, if N is sufficiently large (i.e. if $n=\log_m N$ is sufficiently large), one has $N^{-\Omega(\eta^2/m^2)}<\eta/k$ and is thus absorbed in the $O(\eta)$ error term. ■

The following lemma is the source of hardness of our input instance:

Lemma 12 *For every $k\geq 2$, every $\eta\in(0,1/2)$, every integer m a multiple of k, every $(j_1,\ldots,j_k)\in B_1\times\ldots\times B_k$, for every $i=0,\ldots,k-1$ for every edge (x,y), $x\in S_i^*$ either $y\in T_k$ or $(x,y)\in E_i^{j_{i+1}}$.*

Proof: Consider a edge (x,y) with $x\in S_i^*$ that is not in $E_i^{j_{i+1}}$. We now show that $y\in T_k$, proving the lemma. Let $j\neq j_{i+1}\in B_{i+1}$ be such that $(x,y)\in E_i^j$ – such a j exists by definition of the edge set E_i (recall (3)). This in particular means that $j_r\neq j$ for all $r=1,\ldots,k$, since $j\in B_{i+1}$ and the blocks $B_r, r=1,\ldots,k$ are disjoint. By definition of E_i^j we have $y\in T_i\setminus T_i^j$ and $x\in S_i^j$. Furthermore, we have $(x-y)_s=0$ for all $s\neq j$. We thus have $x_{j_{i+1}}=y_{j_{i+1}}$ for all $i=0,\ldots,k-1$. But since $x_{j_{i+1}}>\frac{m}{k}$ for all $i=0,\ldots,k-1$ (by definition of S_i^* in (2) and assumption that $x\in S_i^*$), this implies $y_{j_{i+1}}>\frac{m}{k}$ for all $i=0,\ldots,k-1$, so $y\in T_k$ (by definition of T_k, see (1)).

■

Lemma 13 *For every $k\geq 2$, $\eta\in(0,1/2)$, if m is an integer multiple of k such that $m\geq C\eta^{-2}k\log\eta^{-1}$ for a sufficiently large constant $C>0$, and if the input graph $G=G(j_1,\ldots,j_r)$ is selected according to the input distribution $\mathcal{D}$ defined above, the following conditions hold. If the streaming algorithm, after being presented with edges revealed in the i-th stage for $i=0,\ldots,k-1$, must store a number of edges after each phase, with the overall set of edges remembered over all stages denoted by E', then any matching M_{ALG} contained in E' satisfies*

$$|M_{ALG}|\leq\left(1-(1-1/k)^k\right)|T|+\sum_{i=0}^{k-1}|E_i^{j_{i+1}}\cap E'|+O(\eta)|T|.$$

Proof: Let the constant $C>0$ be sufficiently large as dictated by Lemmas 8 and 10. We consider the standard reduction of bipartite matching to max-flow (i.e. connect source s to S, sink t to T) and exhibit a cut in the graph $(S\cup\{s\},T\cup\{t\},E')$ of value at most $\left(1-(1-1/k)^k\right)|T|+\sum_{i=0}^{k-1}|E_i^{j_{i+1}}\cap E'|+O(\eta)|T|$. By max-flow/min-cut theorem this gives the result.

We now exhibit a cut in this graph and upper bound its size. The source side of the cut is $\{s\}\cup S_k\cup T_k\cup\bigcup_{i=0}^{k-1}S_i^*$. By Lemma 12 edges incident on $S_i^*, i=0,\ldots,k-1$ either belong to the matching $M_i^{j_{i+1}}$ or go to

T_k, so edges incident on $S_i^*, i = 0, \ldots, k-1$ contribute at most $\sum_{i=0}^{k-1} |E_i^{j_{i+1}} \cap E'|$ to the cut value. We thus have that the value of the cut is bounded by

$$|T_k| + \sum_{i=0}^{k-1} |S_i \setminus S_i^*| + \sum_{i=0}^{k-1} |E_i^{j_{i+1}} \cap E'|. \tag{7}$$

It remains to bound the size of T_k, as well as the sizes of $S_i \setminus S_i^*$. We condition on the event $\mathcal{E}_{set-sizes}$ and Lemma 6. Conditioned on this event we have $|T_k| = (1-1/k)^k |T|$ and

$$|S_i^*| = (1 \pm \eta)|T_k|/k = (1 + O(\eta))(1-1/k)^k/k.$$

Similarly, we have by Lemma 6, **(1)** that $|T_i| = (1-1/k)^i |T|$, and thus by Lemma 6, **(2)** that $|S_i| = (1 \pm O(\eta))(1-1/k)^i |T|/k$. Using these bounds we get

$$\begin{aligned}
\sum_{i=0}^{k-1} |S_i \setminus S_i^*| &= \sum_{i=0}^{k-1} |S_i| - \sum_{i=0}^{k-1} |S_i^*| \\
&= (1 \pm O(\eta)) \sum_{i=0}^{k-1} (1-1/k)^i |T|/k - (1 \pm O(\eta))(1-1/k)^k |T| \\
&= (1 \pm O(\eta))(1 - (1-1/k)^k)|T| - (1 \pm O(\eta))(1-1/k)^k |T| \quad \text{(by summing the geometric series)} \\
&= (1 \pm O(\eta))(1 - 2(1-1/k)^k)|T|
\end{aligned}$$

Putting the bounds above together with (7), we thus have that the size of the cut is bounded by

$$\begin{aligned}
&|T_k| + \sum_{i=0}^{k-1} |S_i \setminus S_i^*| + \sum_{i=0}^{k-1} |E_i^{j_{i+1}} \cap E'| \\
&= (1-1/k)^k |T| + (1 \pm O(\eta))(1 - 2(1-1/k)^k)|T| + \sum_{i=0}^{k-1} |E_i^{j_{i+1}} \cap E'| \\
&= (1 \pm O(\eta))(1 - (1-1/k)^k)|T| + \sum_{i=0}^{k-1} |E_i^{j_{i+1}} \cap E'| \\
&= \left(1 - (1-1/k)^k\right)|T| + \sum_{i=0}^{k-1} |E_i^{j_{i+1}} \cap E'| + O(\eta)|T|
\end{aligned}$$

as required. ■

We now prove

Theorem 14 *For every $k \geq 2$, for any $\eta \in (0, 1/k^3)$, if $m \geq C\eta^{-2} k \log \eta^{-1}$ for a sufficiently large absolute constant $C > 0$ is a multiple of k, then if the graph $G = G(j_1, \ldots, j_k)$ is selected according to the input distribution $\mathcal{D}$ defined above, and the algorithm, after being presented with edges revealed in the i-th stage, stores $s = o(\log N)$ edges, the following conditions hold. If M_{ALG} is the maximum matching in the set of edges E' that the algorithm stored over all the stages, one has*

$$|M_{ALG}| \leq \left(1 - (1-1/k)^k\right)|T| + O(\eta)|T|$$

with probability at least $99/100$.

Proof: Denote the set of edges that the algorithm commits to after seeing the subgraph $S_i \times T_i$ by $\tilde{E}_i$. By Lemma 13 the size of the matching that the algorithm outputs at the end is upper bounded by

$$(1-1/k)^k|T| + \sum_{i=0}^{k-1} |E_i^j \cap \widetilde{E}_i|.$$

We will show that with high probability $\sum_{i=0}^{k-1} |E_i^j \cap \widetilde{E}_i| \leq \sum_{i=0}^{k-1} s/|B_{i+1}| = O(k^2 \cdot s/n)$, where s is the number of edges that the algorithm stores at every step. Recall that for each $i = 0, \ldots, k-1$, conditioned on $(j_1, \ldots, j_i)$, the special index j_{i+1} is chosen uniformly at random in B_{i+1}, implying that

$$\mathbb{E}_{j_{i+1}}\left[\left|E_i^{j_{i+1}} \cap \widetilde{E}_i\right| \middle| j_1, \ldots, j_i\right] = \frac{1}{|B_{i+1}|} \sum_{j \in B_{i+1}} |E_i^j \cap \widetilde{E}_i| = \frac{|\widetilde{E}_i|}{|B_{i+1}|} = s/|B_{i+1}|.$$

Summing over all $i = 0, \ldots, k-1$, we get

$$\mathbb{E}\left[\sum_{i=0}^{k-1} |E_i^j \cap \widetilde{E}_i|\right] = \sum_{i=0}^{k-1} s/|B_{i+1}| = O(k^2 \cdot s/n) = o(k^2|T|) = o(|T|),$$

since $s = o(\log N)$ by assumption of the theorem and $\log N = \Theta(n)$ (as m is a constant). The result now follows by Markov's inequality. ■

Theorem 15 *For every $k \geq 2$, every $\delta \in (0,1)$, there exists an input distribution $\mathcal{D}$ on bipartite graphs such that any streaming algorithm that stores $s = o(N \log N)$ edges achieves an approximation ratio of at most $1-(1-1/k)^k + \delta$.*

Proof: Consider the distribution $\mathcal{D}$ with $\eta = c\delta/k^3$ for a sufficiently small constant $c > 0$ and $m \geq C\eta^{-2}k\log\eta^{-1}$ a multiple of k for the constant $C > 0$ from Lemma 8. Then by Theorem 14 one has

$$|M_{ALG}| \leq \left(1-(1-1/k)^k\right)|T| + O(\eta)|T| \leq \left(1-(1-1/k)^k\right)|T| + (\delta/2)|T| = \left(1-(1-1/k)^k + \delta/2\right)N$$

with probability at least $99/100$. At the same time by Lemma 11, conditioned on conditioned on $\mathcal{E}_{large-lines}$ (defined in Lemma 8) and $\mathcal{E}_{set-sizes}$ (defined in Lemma 6), every $(j_1, \ldots, j_k) \in B_1 \times \ldots \times B_k$ the graph $G = G(j_1, \ldots, j_k)$ contains a matching of size at least $(1-O(\eta+\eta^2 k))|S| \geq (1-\delta/10)N$ if N is sufficiently large. Thus, the approximation ratio achieved by the algorithm is at most

$$\frac{1-(1-1/k)^k + \delta/2}{1-\delta/10} \leq 1-(1-1/k)^k + \delta,$$

as required.

It remains to note that by Lemma 8 the event $\mathcal{E}_{large-lines}$ occurs with probability at least $1-k(\log N)^k e^{-\Omega(\eta N/k)} \geq 1-o(1)$ over the choice of Z_x's since k and η are independent of N by our setting of parameters. Similarly, by Lemma 6 the event $\mathcal{E}_{set-sizes}$ occurs with probability at least $1-k(\log N)^k e^{-\Omega(\eta N/k)} \geq 1-o(1)$. Thus, the upper bound on the approximation ratio achieve by the algorithm holds with probability at least $99/100 - o(1) \geq 98/100$, as required.

■

3 Single pass streaming lower bound

In the rest of the section we define a distribution on input instances for our problem of approximating maximum matchings in a single pass in the streaming model. Our construction follows its simple version presented in Section 2. A major difference is that we replace coordinate directions with an exponential size family of nearly orthogonal vectors, thereby achieving a lower bound of $n^{1+\Omega(1/\log\log n)}$ on the space complexity of obtaining a better than $1-1/e$ approximation in a single pass. This approach is inspired by techniques for constructing Ruzsa-Szemerédi graphs pioneered in [FLN$^+$02] and extensions developed in [GKK12].

3.1 Construction of host graphs $G(\mathbf{u}_1,\ldots,\mathbf{u}_k)$

We first introduce notation. Each graph in our family of host graphs will be indexed by a k-tuple of vectors $(\mathbf{u}_1,\ldots,\mathbf{u}_k)\in\mathcal{F}_1\times\ldots\times\mathcal{F}_k$, where $\mathcal{F}_j, j=1,\ldots,k$ are families of vectors in $\{0,1\}^m$. We choose $\mathcal{F}_j$ so that vectors in $\bigcup_{j=1}^k\mathcal{F}_j =: \mathcal{F}$ are of equal Hamming weight and nearly orthogonal. Specifically, the following lemma guarantees the existence of a large family $\mathcal{F}$ such that for every $\mathbf{u},\mathbf{v}\in\mathcal{F}$ it holds that $|\mathbf{u}|=|\mathbf{v}|=w$ and $(\mathbf{u},\mathbf{v})\le\epsilon w$, where $w=\Theta(\epsilon^2 m)$. Here for a vector $\mathbf{u}\in\{0,1\}^m$ we write $|\mathbf{u}|$ to denote the Hamming weight of $\mathbf{u}$. We assume from now on that $|\mathcal{F}_1|=|\mathcal{F}_2|=\ldots=|\mathcal{F}_k|=d$ for a parameter d. The lemma below shows that we can have $d=2^{\Omega(\epsilon^2 m)}$. The specific form of the dependence of the exponent on ϵ will not be important for the qualitative nature of our results, however, as we will ultimately set ϵ to be a small constant.

We will use the following standard lower bound on the size of such families:

Lemma 16 *For any $\epsilon\in(0,1)$, any integers $m\ge 1$ and $w=(\epsilon/2)m$, there exists a collection $\mathcal{F}_{m,w,\epsilon}\subset\{0,1\}^m$ of vectors of Hamming weight w with $\log|\mathcal{F}_{m,w,\epsilon}|=\Omega(\epsilon^2 m)$ such that for all $\mathbf{u}\neq\mathbf{u}'\in\mathcal{F}_{w,\epsilon}$, $(\mathbf{u},\mathbf{u}')<\epsilon w$.*

Proof: The proof is via the probabilistic method. Partition $[m]$ into w subsets $I_1,\ldots,I_w$, with $|I_s|=m/w$ for $s=1,\ldots,w$. We pick $\mathbf{u}_1,\ldots,\mathbf{u}_N$ independently as follows. For every $j=1,\ldots,N$, the vector $\mathbf{u}_j$ includes exactly one random element of I_s for each $s=1,\ldots,w$. This ensures that the Hamming weight of each $\mathbf{u}_j$ is exactly w.

We now show that the vectors have small intersection size with high probability. Fix $i\neq j\in[N]$. Imagine $\mathbf{u}_i$ being fixed and picking the w elements of $\mathbf{u}_j$ one by one. Let X_s denote the indicator random variable for the event that the sth element of $\mathbf{u}_j$ (picked from I_s) is also in S_i. Then $(\mathbf{u}_i,\mathbf{u}_j)=\sum_{s=1}^w X_k$, and we set $\mu:=\mathbb{E}[(\mathbf{u}_i,\mathbf{u}_j)]$. Note that $\mu=(w/m)\cdot w$, since for every $s=1,\ldots,w$ the vector $\mathbf{u}_i$ has exactly one nonzero coordinate in I_s, and the probability that $\mathbf{u}_j$ chooses the same coordinate is $1/|I_s|=w/m$. We have $\mathbb{P}[(\mathbf{u}_i,\mathbf{u}_j)\ge\epsilon w]=\mathbb{P}[\sum_{s=1}^w X_s\ge 2\mu]$ The random variables X_s are independent and thus the Chernoff bound yields

$$\mathbb{P}[(\mathbf{u}_i,\mathbf{u}_j)\ge 2\mu)\le\left(\frac{e}{4}\right)^{\mu}\le e^{-\Omega((w/m)w)}\le e^{-c\epsilon^2 m}$$

for a constant $c>0$. Setting $N=2^{(\ln_2 e)c\epsilon^2 m/2}$ so that $\binom{N}{2}<N^2=2^{(\ln_2 e)c\epsilon^2 m}=e^{c\epsilon^2 m}$, by a union bound with positive probability $|\mathbf{u}_i\cap\mathbf{u}_j|<\epsilon w$ for all $i\neq j$, simultaneously, as desired. Note for this choice of N, we have $\log|\mathcal{F}_{m,w,\epsilon}|=\log N=\Theta(\epsilon^2 m)$. ∎

We also associate with each $\mathbf{u}\in\mathcal{F}_j, j=1,\ldots,k$ a random variable $U_{\mathbf{u}}$ that is uniformly distributed over the integers

$$\{0,1,\ldots,k/\theta-1\}\cdot W\cdot(\theta/k), \tag{8}$$

where $\theta\in(0,1)$ is a parameter that we will set to a small constant times $1/\text{poly}(k)$, and W is a parameter that will later set to $\text{poly}(k)\cdot w$ (where w is the Hamming weight of the vectors in the collection $\mathcal{F}$). The variables $U_{\mathbf{u}}$ and $U_{\mathbf{v}}$ are independent for $\mathbf{u}\neq\mathbf{v}$.

As before, the sides of the bipartition of the graph $G(\mathbf{u}_1, \ldots, \mathbf{u}_k)$ that we need to construct are denoted by T and $S = S_0 \cup \ldots \cup S_k$, where $S_0 \cup S_1 \cup \ldots \cup S_k$ is a partition of S. We use the notation $[a] = \{1, \ldots, a\}$ for integer $a \geq 1$. In our construction the $T = T^0$ side of the graph is identified with a hypercube $[m^4]^m$ for a value of m to be chosen later, and each set $S_i, i = 0, \ldots, k-1$ is identified with a subsampled hypercube $[m^4]^m$. The vertices of the last set S_k do not have any special structure. Vertices $x \in T$ or $y \in S_i$ will often be treated as points $x, y \in [m^4]^m$. For $x \in T$ and $\mathbf{u} \in \mathcal{F}$ we use the dot product notation $(x, \mathbf{u}) = \sum_{i=1}^{m} x_i \cdot \mathbf{u}_i \in \mathbb{Z}$. For an interval $[a, b]$ and a number W we will write $[a, b] \cdot W$ to denote the set of integers belonging to the interval $[a \cdot W, b \cdot W]$. Finally, for an integer i and an integer W we will write $i \mod W$ to denote the residue of i modulo W that belongs to $[0, W-1]$.

3.1.1 Defining the vertex set of the graph $G(\mathbf{u}_1, \ldots, \mathbf{u}_k)$

We will use

Definition 17 (Ground sets X, Y) *Let $X^* = Y = [m^4]^m$ for some integer $m > 0$. Let X be a random subset of X^* where each point of X^* appears independently with probability $1/k$.*

We will refer to vertices in X and Y as points in $[m^4]^m$. The host graph G is generated by first selecting a k-tuple $(\mathbf{u}_1, \ldots, \mathbf{u}_k) \in \mathcal{F}_1 \times \ldots \times \mathcal{F}_k$, and then defining the vertex and edge set as we describe below. Before proceeding with the construction, we list relevant parameters here.

Parameters of the construction

- k – the number of phases in the hard input distribution;
- w — Hamming weight of binary vectors in $\mathcal{F}$;
- ϵ – upper bound on the maximum dot product of any pair of distinct vectors from $\mathcal{F}$, normalized by their Hamming weight w;
- η – small constant governing separation of red and blue vertices – see (9) and (10).

Sets of red, white and blue vertices R, W, B. Consider fixed $\mathbf{w} \in \{0,1\}^m$, and let

$$\begin{aligned} R^Y(\mathbf{w}) &= \{y \in Y : ((y, \mathbf{w}) + U(\mathbf{w})) \mod W \in [0, 1/k) \cdot W\} \\ &\quad \text{(red vertices with respect to } \mathbf{w}) \\ W^Y(\mathbf{w}) &= \{y \in Y : ((y, \mathbf{w}) + U(\mathbf{w})) \mod W \in ([1/k, 1/k+\eta) \cup [1-\eta, 1)) \cdot W\} \\ &\quad \text{(white vertices with respect to } \mathbf{w}) \\ B^Y(\mathbf{w}) &= \{y \in Y : ((y, \mathbf{w}) + U(\mathbf{w})) \mod W \in [1/k+\eta, 1-\eta) \cdot W\} \\ &\quad \text{(blue vertices with respect to } \mathbf{w}) \end{aligned} \tag{9}$$

It is convenient to also define

$$\begin{aligned} R^{X^*}(\mathbf{w}) &= \{x \in X^* : ((x, \mathbf{w}) + U(\mathbf{w})) \mod W \in [0, 1/k) \cdot W\} \\ W^{X^*}(\mathbf{w}) &= \{x \in X^* : ((x, \mathbf{w}) + U(\mathbf{w})) \mod W \in ([1/k, 1/k+\eta) \cup [1-\eta, 1)) \cdot W\} \\ B^{X^*}(\mathbf{w}) &= \{x \in X^* : ((x, \mathbf{w}) + U(\mathbf{w})) \mod W \in [1/k+\eta, 1-\eta) \cdot W\}, \end{aligned} \tag{10}$$

as well as let

$$R^X(\mathbf{w}) = R^{X^*}(\mathbf{w}) \cap X, \;\; W^X(\mathbf{w}) = W^{X^*}(\mathbf{w}) \cap X, \quad B^X(\mathbf{w}) = B^{X^*}(\mathbf{w}) \cap X. \tag{11}$$

The intuition for these sets is simple: ideally, we would like to partition vertices $y \in T_i$ into two classes, depending on whether their dot product with $\mathbf{w}$ is in $[0, 1/k) \cdot W$ (red points) or $[1/k, 1) \cdot W$(blue points; we ignore the shift $U(\mathbf{w})$ for this intuitive discussion), and then match points in one color class in X to points in the other color class in Y. This would work fine if the set $\mathcal{F}$ of vectors that we use contained orthogonal vectors only, as in our toy construction in Section 2. Since the family of vectors $\mathcal{F}$ that we use consists of vectors with small (constant) dot products, we need a 'buffer' between the two classes above, provided by the set of white vertices W^Y.

Nested sequence of sets T_i and sets S_i. For all $i = 0, \ldots, k$ let

$$\begin{aligned} T_i &= \{y \in Y : ((y, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in [1/k, 1) \cdot W, \text{ for all } j \in [1 : i]\} \\ S_i &= \{x \in X : ((x, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in [1/k, 1) \cdot W, \text{ for all } j = [1 : i]\}, \end{aligned} \tag{12}$$

so that $T_0 = Y$ and $S_0 = X$. For every $i = 0, \ldots, k-1$ and $\mathbf{w} \in \mathcal{F}_{i+1}$ we let

$$\begin{aligned} T_i^{\mathbf{w}} = \{y \in Y : &((y, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in [1/k, 1) \cdot W, \text{ for all } j \in [1 : i] \\ &\text{and} \\ &((y, \mathbf{w}) + U(\mathbf{w})) \mod W \in [1/k, 1) \cdot W\} \\ S_i^{\mathbf{w}} = \{x \in X : &((x, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in [1/k, 1) \cdot W, \text{ for all } j = [1 : i] \\ &\text{and} \\ &((x, \mathbf{w}) + U(\mathbf{w})) \mod W \in [1/k, 1) \cdot W\} \end{aligned} \tag{13}$$

Also, let

$$\begin{aligned} S_i^* = \{x \in S_i : &((x, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in [1/k, 1) \cdot W, \\ &\text{for all } j \in [1 : k]\}. \end{aligned} \tag{14}$$

Note that the sets S_i^* are obtained from S_i by adding extra constraints on dot products with vectors $\mathbf{u}_j$, namely for $j = i+1, \ldots, k$ (this is because all vertices in S_i already satisfy the constraints above for $j = 1 : i$ by definition of S_i).

Vertex set of $G(\mathbf{u}_1, \ldots, \mathbf{u}_k)$. The graph $G(\mathbf{u}_1, \ldots, \mathbf{u}_k)$ whose edges we define shortly will be a bipartite graph with the sides of the bipartition given by $T = T_0$ and $S = S_0 \cup \ldots \cup S_k$, where T_0 and $S_i, i = 0, 1, \ldots, k$ are as defined above. Note that the union of S_i's in the definition of S is understood as a disjoint union. In other words, vertices in both S and T are naturally labelled with points on the hypercube in $[m^4]^m$. These labels are distinct for vertices in T, but not for vertices in S. However, such labels are distinct for vertices in S_i for every $i = 0, 1, \ldots, k$. We denote the number of vertices on the Q side of the bipartition, i.e., in T, by n. We will have $|S| = O(|T|)$, so that the total number of vertices in our instance is $O(n)$.

Estimates on the size of $T_i, S_i, T_i^{\mathbf{w}}, S_i^{\mathbf{w}}, R, B, W$. We will need the following lemma, whose proof is given in Appendix A

Lemma 18 *For every $m \geq 2$, integer $W \geq 1$ and $\delta' \in (0, 1)$ such that $1/\delta'$ is an integer, if $Y = [m^4]^m$ and the set $\mathcal{S}$ is defined by*

$$\mathcal{S} = \{y \in Y : (y, \mathbf{u}) + \Delta_{\mathbf{u}} \mod W \in [a_{\mathbf{u}}, b_{\mathbf{u}}) \cdot W, \textit{ for all } \mathbf{u} \in \mathcal{U}\},$$

where $\mathcal{U}$ is a collection of binary vectors of fixed length w and $a_{\mathbf{u}}, b_{\mathbf{u}} \in [0,1]$ are constant integer multiples of $1/L$ for an integer L, the following conditions hold if W is an integer multiple of $w \cdot lcm(L, 1/\delta')$, $\Delta_{\mathbf{u}}/W$ are multiples of $1/L$ and m is sufficiently large.

If $\max_{\substack{\mathbf{u}\in\mathcal{U},\mathbf{v}\in\mathcal{U}\\ \mathbf{u}\neq\mathbf{v}}}(\mathbf{u},\mathbf{v})/|\mathbf{v}| \leq \delta'$*, then*

$$\left||\mathcal{S}| - |Y| \cdot \prod_{\mathbf{u}\in\mathcal{U}} (b_{\mathbf{u}} - a_{\mathbf{u}})\right| \leq |\mathcal{U}|^2(6L\delta' + 4/m) \cdot |Y|.$$

We now apply Lemma 18 to bound the size of various relevant subsets of Y and X. We gather the resulting bound in the following

Lemma 19 *There exists an event $\mathcal{E}$ over X that occurs with probability at least $99/100$ such that the following bounds hold conditioned on $\mathcal{E}$.*

For every $\epsilon \in (0,1)$, every integer $k \geq 2$, every choice of shifts $U(\mathbf{u}) \in \mathcal{F}$, every $(\mathbf{u}_1, \ldots, \mathbf{u}_k) \in \mathcal{F}_1 \times \mathcal{F}_2 \times \ldots \times \mathcal{F}_k$, if W (see (12) and (14)) is an integer multiple of $w \cdot k/(\epsilon \cdot \theta)$ (see (8)), then for sufficiently large integer m

(1) $|T_i| = (1-1/k)^i|Y| + \Delta_i$, $|\Delta_i| = O(k^3\epsilon/\theta) \cdot |Y|$ *for every* $i \in \{0,1,2,\ldots,k\}$*;*

(2) $|S_i| = \frac{1}{k}((1-1/k)^i|Y| + \Delta_i$, $|\Delta_i| = O(k^3\epsilon/\theta) \cdot |Y|$ *for every* $i \in \{0,1,2,\ldots,k\}$*;*

(3) $|S_i^*| = \frac{1}{k}(1-1/k)^k|Y| + \Delta_i$, $|\Delta_i| = O(k^3\epsilon/\theta) \cdot |Y|$*;*

(4) $|T_i^{\mathbf{w}}| = (1-1/k)^{i+1}|Y| + \Delta_i$, $|\Delta_i| = O(k^3\epsilon/\theta) \cdot |Y|$ *for every* $i \in \{0,1,2,\ldots,k-1\}$ *and* $\mathbf{w} \in \mathcal{F}_{i+1}$*;*

(5) $|S_i^{\mathbf{w}}| = \frac{1}{k}(1-1/k)^{i+1}|Y| + \Delta_i$, $|\Delta_i| = O(k^3\epsilon/\theta) \cdot |Y|$ *for every* $i \in \{0,1,2,\ldots,k-1\}$ *and* $\mathbf{w} \in \mathcal{F}_{i+1}$.

Proof: We fix the values of the shifts $U(\mathbf{u}), \mathbf{u} \in \mathcal{F}$ as well as the sequence $(\mathbf{u}_1, \ldots, \mathbf{u}_k) \in \mathcal{F}_1 \times \mathcal{F}_2 \times \ldots \times \mathcal{F}_k$, and take a union bound over such fixings later (this is important for establishing the bounds on various subsets of S, as those depend on the random choice of X; see (12), (13) and (14)). By (12) we have

$$\begin{aligned} T_i &= \{y \in Y : ((y, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in [1/k, 1) \cdot W, \text{ for all } j \in [1:i]\} \\ S_i &= \{x \in X : ((x, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in [1/k, 1) \cdot W, \text{ for all } j = [1:i]\}. \end{aligned}$$

We start with T_i, where we apply Lemma 18 with $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_i\}$. We thus have $a_{\mathbf{u}} = \{1/k + U(\mathbf{u})/W\}$, $b_{\mathbf{u}} = \{1 + U(\mathbf{u})/W\}$ for all $\mathbf{u} \in \mathcal{U}$ (where $\{\cdot\}$ stands for the fractional part of the argument). Since $U(\mathbf{w})$ are integer multiples of $\theta W/k$ by definition (see (8)), we get that setting $L = k/\theta$ ensures that $a_{\mathbf{u}}, b_{\mathbf{u}}$ are multiples of $1/L$. Recall that vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k \in \{0,1\}^m$ have Hamming weight w and $\max_{s\neq t}(\mathbf{u}_s, \mathbf{u}_t)/w \leq \epsilon$ by assumption of the lemma. Since further W is an integer multiple of $w \cdot k/(\epsilon \cdot \theta)$, we get that indeed W is an integer multiple of $w \cdot \text{lcm}(L, 1/\epsilon)$, and hence the preconditions of Lemma 18 are satisfied. We now get by Lemma 18, using the fact that $\prod_{\mathbf{u}\in\mathcal{U}}(b_{\mathbf{u}} - a_{\mathbf{u}}) = (1-1/k)^i$ by our setting of parameters, that

$$\begin{aligned} \left||T_i| - (1-1/k)^i|Y|\right| &\leq |\mathcal{U}|^2(6L\epsilon + 4/m) \cdot |Y| \\ &\leq k^2(6(k/\theta)\epsilon + 4/m) \cdot |Y| \\ &= O(k^3\epsilon/\theta) \cdot |Y|, \end{aligned}$$

where in the last step we used the assumption of our lemma that m if sufficiently large as a function of k and ϵ. This proves **(1)**. The proof of **(4)** is similar, with $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_i, \mathbf{w}\}$.

Similarly, since every element of $X^* = Y$ appears in X independently with probability $1/k$ (see Definition 17), an application of Lemma 18 as above shows that

$$\mathbb{E}_X[|S_i|] = \frac{1}{k}((1-1/k)^i|Y| \pm O(k^3\epsilon/\theta)\cdot|Y|).$$

We thus also get by an application of Chernoff bounds (Theorem 5) we get that for every $i = 0,\ldots,k$

$$\mathbb{P}\left[||S_i| - \mathbb{E}[|S_i|]| > \epsilon\cdot|Y|\right] < 2e^{-\epsilon^2|Y|/(3k)}. \tag{15}$$

The bound above is for a fixed choice of the shifts $U(\mathbf{u}), \mathbf{u}\in\mathcal{F}$. The number of such choices is bounded by $(m^5)^{2^m} \leq e^{|Y|^{1/2}}$ when m is larger than a constant. The number of choice of $(\mathbf{u}_1,\ldots,\mathbf{u}_k)\in\mathcal{F}_1\times\mathcal{F}_2\times\ldots\times\mathcal{F}_k$ is bounded by $2^{k2^m} \leq e^{|Y|^{1/2}}$ as well, and therefore we have $||S_i| - \mathbb{E}[|S_i|]| \leq \epsilon\cdot|Y|$ for every choice of shifts with probability at least $1-2e^{-\epsilon^2|Y|/(6k)}$. Denote the success event by $\mathcal{E}_i$. Conditioned on $\bigcap_{i=0}^{k-1}\mathcal{E}_i$ one has $\left||S_i| - \frac{1}{k}((1-1/k)^i|Y|\right| = O(k^2\epsilon/\theta)\cdot|Y|$, proving **(2)**.

Similarly, since every element of $X^* = Y$ appears in X independently with probability $1/k$ (see Definition 17), an application of Lemma 18 as above shows that for every $\mathbf{w}\in\mathcal{F}_{i+1}$

$$\mathbb{E}_X[|S_i^{\mathbf{w}}|] = \frac{1}{k}((1-1/k)^{i+1}|Y| \pm O(k^3\epsilon/\theta)\cdot|Y|).$$

We thus also get by an application of Chernoff bounds (Theorem 5) we get that for every $i = 0,\ldots,k-1$ and $\mathbf{w}\in\mathcal{F}_{j+1}$

$$\mathbb{P}\left[||S_i^{\mathbf{w}}| - \mathbb{E}[|S_i^{\mathbf{w}}|]| > \epsilon\cdot|Y|\right] < 2e^{-\epsilon^2|Y|/(3k)}. \tag{16}$$

Similarly to the above, we take a union bound over all fixings of shifts $U(\mathbf{u}), \mathbf{u}\in\mathcal{F}$ and choices of $(\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_k)\in\mathcal{F}_1\times\mathcal{F}_2\times\ldots\times\mathcal{F}_k$, getting an upper bound of $2e^{-\epsilon^2|Y|/(6k)}$ on the probability of the failure event. Denote the success event by $\mathcal{E}_{i,\mathbf{w}}$. Conditioned on $\bigcap_{i=0}^{k-1}\bigcap_{\mathbf{w}\in\mathcal{F}_{i+1}}\mathcal{E}_{i,\mathbf{w}}$ one has $\left||S_i^{\mathbf{w}}| - \frac{1}{k}((1-1/k)^{i+1}|Y|\right| = O(k^3\epsilon/\theta)\cdot|Y|$, proving **(5)**.

Finally, recall that by (14) one has for every $i\in 0,1,\ldots,k-1$

$$\begin{aligned}
S_i^* &= \{x\in S_i : ((x,\mathbf{u}_l)+U(\mathbf{u}_l)) \mod W \in [1/k,1)\cdot W, \text{ for all } l\in[1:k]\}\\
&= \{x\in[m^4]^m\cap X : ((x,\mathbf{u}_l)+U(\mathbf{u}_l)) \mod W \in [1/k,1)\cdot W, \text{ for all } l\in[1:k]\}\\
&= T_k\cap X.
\end{aligned}$$

We now have by **(1)** that

$$\left||T_k| - (1-1/k)^k|Y|\right| = O(k^3\epsilon/\theta)\cdot|Y|.$$

We thus get, since X contains every element of $[m^4]^m$ independently with probability $1/k$ by Definition 17, that $\mathbb{E}_X[|S_i^*|] = |T^k|/k$, and thus by Chernoff bounds (Theorem 5)

$$\mathbb{P}\left[||S_i^*| - \mathbb{E}[|S_i^*|]| > \epsilon\cdot|Y|\right] < 2e^{-\epsilon^2|Y|/(3k)}. \tag{17}$$

Similarly to the above, we take a union bound over all fixings of shifts $U(\mathbf{u}), \mathbf{u}\in\mathcal{F}$ and choices of $(\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_k)\in\mathcal{F}_1\times\mathcal{F}_2\times\ldots\times\mathcal{F}_k$, getting an upper bound of $2e^{-\epsilon^2|Y|/(6k)}$ on the probability of the failure event. Denote the success event by $\mathcal{E}_{i,*}$. Conditioned on $\bigcap_{i=0}^{k-1}\mathcal{E}_{i,*}$ one has $\left||S_i^*| - \frac{1}{k}((1-1/k)^*|Y|\right| = O(k^3\epsilon/\theta)\cdot|Y|$, as required.

We now let $\mathcal{E} = \left(\bigcap_{i=1}^{k} \mathcal{E}_i\right) \cap \left(\bigcap_{i=1}^{k} \mathcal{E}_{i,*}\right) \cap \left(\bigcap_{i=0}^{k-1} \bigcap_{\mathbf{w}\in\mathcal{F}_{i+1}} \mathcal{E}_{i,\mathbf{w}}\right)$. Using a union bound together with (15), (16) and (17) we get

$$\begin{aligned}
\mathbb{P}[\mathcal{E}] &\geq 1 - \sum_{i=0}^{k-1} \mathbb{P}[\bar{\mathcal{E}}_i] - \sum_{i=0}^{k-1} \mathbb{P}[\bar{\mathcal{E}}_{i,*}] - \sum_{i=0}^{k-1} \sum_{\mathbf{w}\in\mathcal{F}_{i+1}} \mathbb{P}[\bar{\mathcal{E}}_{i,\mathbf{w}}] \\
&\geq 1 - 2k2^m 2e^{-\epsilon^2|Y|/(6k)} \\
&\geq 1 - 2k2^m 2e^{-\epsilon^2 m^{4m}/(6k)} \\
&\geq 99/100
\end{aligned}$$

as long as m is sufficiently large as a function of ϵ and k. ■

3.1.2 Defining the edge set of the graph $G(\mathbf{u}_1, \ldots, \mathbf{u}_k)$

First, the only edges incident on vertices in S_k are the edges of a perfect matching between S_k and T^k. In the rest of the section we define edges incident on S_i, $i = 0, \ldots, k-1$. The following definition will be useful in the analysis. Let Bad $\subseteq [m^4]^m$ be defined by

$$\text{Bad} := \{y \in [m^4]^m : \exists i \text{ s.t. } y_i < m^2 \text{ or } y_i > m^4 - m^2\}. \tag{18}$$

Note that $|\text{Bad}| \leq 2m^3/m^4 = o(1)$ by a union bound.

For each $i = 0, \ldots, k-1$ we will have $\Gamma_G(S_i) \subseteq T_i$ (see (12) for the definitions of S_i and T_i). The edges incident to S_i can be partitioned into an induced union of nearly regular constant degree subgraphs. Each such subgraph is indexed by a vector $\mathbf{w} \in \mathcal{F}_{i+1}$, and is denoted by $H_i^{\mathbf{w}}$. We now give the construction of these subgraphs.

The graph $H_i^{\mathbf{w}}$ is a disjoint union of constant size complete bipartite graphs, where each such constant size graph corresponds to a set of points on the integer lattice that lie on a short line segment in direction $\mathbf{w}$ (recall that $\mathbf{w} \in \{0,1\}^m$). In what follows we first define the relevant lines (Sets $\mathcal{L}^Y$ and $\mathcal{L}^X$), and then define the edges of $H_i^{\mathbf{w}}$.

Defining sets of lines $\mathcal{L}^Y$ and $\mathcal{L}^X$. For an arbitrary $y \in R^Y$ let

$$\ell(y) := \left\lfloor \frac{(y, \mathbf{w}) + U(\mathbf{w})}{W} \right\rfloor,$$

and define

$$\text{line}^Y(y, \mathbf{w}) := \left\{y' \in R^Y : y' = y + \lambda \cdot \mathbf{w} \text{ such that } \ell(y') = \ell(y)\right\}. \tag{19}$$

Similarly, for $x \in B^{X^*}$ let

$$\text{line}^{X^*}(x, \mathbf{w}) := \left\{x' \in B^{X^*} : x' = x + \lambda \cdot \mathbf{w} \text{ such that } \ell(x') = \ell(x)\right\}.$$

and for $x \in B^X$ let

$$\text{line}^X(x, \mathbf{w}) := \left\{x' \in B^X : x' = x + \lambda \cdot \mathbf{w} \text{ such that } \ell(x') = \ell(x)\right\}. \tag{20}$$

Note that for every fixed $\mathbf{w}$ and every pair $x_1, x_2 \in X \setminus \text{Bad}$ one has either $\text{line}^{X^*}(x_1, \mathbf{w}) = \text{line}^{X^*}(x_2, \mathbf{w})$ or $\text{line}^{X^*}(x_1, \mathbf{w}) \cap \text{line}^{X^*}(x_2, \mathbf{w}) = \emptyset$. Analogous properties hold for line^X and line^Y. Let

$$\mathcal{L}^X(\mathbf{w}) = \bigcup_{x\in X\setminus\text{Bad}} \text{line}^X(x, \mathbf{w}) \text{ and } \mathcal{L}^Y(\mathbf{w}) = \bigcup_{y\in Y\setminus\text{Bad}} \text{line}^Y(y, \mathbf{w}).$$

Note that for every $y \in Y \setminus \text{Bad}$ and every $\mathbf{w}$ there exists a unique line $L^X \in \mathcal{L}^X(\mathbf{w})$ such that for every $x \in L^X$ and $y \in \text{line}(y, \mathbf{w}) =: L^Y$ one has $x - y = \lambda \mathbf{w}$ for some integer λ and $\ell(x) = \ell(y)$. We call L^X the pair of L^Y. We denote the function mapping y-lines to their corresponding pair x-lines by $\pi_{\mathbf{w}} : \mathcal{L}^Y(\mathbf{w}) \to \mathcal{L}^X(\mathbf{w})$. Let $\mathcal{L}^{X^*}(\mathbf{w})$ and $\pi^*_{\mathbf{w}}$ be defined analogously.

We now give bounds on the size of lines. We start with

Claim 20 (Size of line$^Y(y, \mathbf{w})$) *If $m \geq W/|\mathbf{w}|$ and $W/|\mathbf{w}|$ is an integer multiple of k, then for all $y \in R^Y \setminus$ Bad, $\mathbf{w} \in \bigcup_{i=1}^k \mathcal{F}_i$ one has $|\text{line}^Y(y, \mathbf{w})| = W/(k|\mathbf{w}|)$.*

Proof: First note that for every integer λ, $\mathbf{w} \in \bigcup_{i=1}^k \mathcal{F}_i$ and every $y \in Y$ one has, letting $y' = y + \lambda \mathbf{w}$,

$$(y', \mathbf{w}) + U(\mathbf{w}) = ((y + \lambda \cdot \mathbf{w}, \mathbf{w}) + U(\mathbf{w})) = ((y, \mathbf{w}) + U(\mathbf{w})) + \lambda |\mathbf{w}|.$$

Note that every λ that results in $\ell(y + \lambda \mathbf{w}) = \ell(y)$ satisfies $|\lambda| \leq W/|\mathbf{w}|$, and thus by our assumption $y \in Y \setminus$ Bad we have $y + \lambda \mathbf{w} \in Y$ since $m \geq W/|\mathbf{w}|$ by assumption of the claim.

Recall that $y' \in \text{line}^Y(y, \mathbf{w})$ amounts to two conditions: $y' \in R^Y$ and $\ell(y') = \ell(y)$, where the former constraint is

$$((y', \mathbf{w}) + U(\mathbf{w})) \pmod W \in [0, 1/k) \cdot W. \tag{21}$$

Letting $z := \left\lfloor \frac{((y, \mathbf{w}) + U(\mathbf{w})) \bmod W}{|\mathbf{w}|} \right\rfloor$, we note that the set of values of λ results in (21) being satisfied at the same time as $\ell(y') = \ell(y)$ is exactly $\{-z, -z+1, \ldots, -z + W/(k|\mathbf{w}|) - 1\}$. We thus have that $|\text{line}^Y(y, \mathbf{w})| = W/(k|\mathbf{w}|)$ for all $y \in Y \setminus$ Bad, as required. ■

We have

Claim 21 (Size of line$^X(x, \mathbf{w})$) *For every $\epsilon \in (0, 1)$, every $\eta \in (0, 1/10)$ such that $1/\eta$ is an integer, if $m \geq W/|\mathbf{w}|$ is sufficiently large and $W/|\mathbf{w}|$ is an integer multiple of lcm$(k, 1/\eta)$, $W/|\mathbf{w}| \geq \frac{48k^2 \ln(1/\eta)}{\epsilon^2}$, the following conditions hold.*

With probability at least $99/100$ over the choice of $X \subseteq X^*$ for every setting of $U(\mathbf{w})$, $\mathbf{w} \in \bigcup_{i=1}^k \mathcal{F}_i$, for all but $2\eta^2|Y|$ points $x \in B^X \setminus$ Bad, every $\mathbf{w} \in \bigcup_{i=1}^k \mathcal{F}_i$ one has $|\text{line}^X(x, \mathbf{w})| \geq \frac{1}{k}(1-1/k)(1-4\eta-\epsilon)W/|\mathbf{w}|$ for sufficiently large m as a function of $\eta, k, W/|\mathbf{w}|$ and ϵ.

Proof: For every integer λ and every $x \in X^* \setminus$ Bad one has

$$((x + \lambda \cdot \mathbf{w}, \mathbf{w}) + U(\mathbf{w})) = ((x, \mathbf{w}) + U(\mathbf{w})) + \lambda |\mathbf{w}|.$$

Letting $z := \left\lfloor \frac{((y, \mathbf{w}) + U(\mathbf{w})) \bmod W}{|\mathbf{w}|} \right\rfloor$, we note that the set of values of λ results in the equation above being satisfied at the same time as $\ell(y') = \ell(y)$ is exactly $\{-z + \frac{W}{|\mathbf{w}|}\left(\frac{1}{k} + \eta\right), \ldots, -z + \frac{W}{|\mathbf{w}|}(1 - \eta) - 1\}$. We thus get that the set of values of λ that result in $x' = x + \lambda \cdot \mathbf{w} \in \text{line}^{X^*}(x, \mathbf{w})$ has size $(1 - 1/k - 2\eta)W/|\mathbf{w}|$, as required.

Note that every λ that results in $\ell(x + \lambda \mathbf{w}) = \ell(x)$ satisfies $|\lambda| \leq W/|\mathbf{w}|$, and thus by our assumption $x \in X^* \setminus$ Bad we have $x' + \lambda \mathbf{w} \in X^*$ since $m \geq W/|\mathbf{w}|$ by assumption of the claim. In order to establish the claim, it suffices to analyze the sampling process involved in constructing X from X^*.

Since for every L^{X^*} the set L^X is a random subsampling of L^{X^*}, where each element of X^* is included in X independently with probability $1/k$, we have $\mathbb{E}\left[|L^X|\right] = \frac{1}{k}|L^{X^*}| = \frac{1}{k}(1 - 1/k - 2\eta)W/|\mathbf{w}|$ for every $y \in Y \setminus$ Bad. Since X is obtained from X^* by independent sampling at rate $1/k$, we get by the Chernoff bound (Theorem 5)

$$\mathbb{P}\left[|L^X| \notin (1 \pm \epsilon)\frac{1}{k}\left(1 - 1/k - 2\eta\right)|L^{X^*}|\right] \leq 2e^{-\epsilon^2 |L^{X^*}|/(12k)},$$

where we used the fact that $1 - 1/k - 2\eta \geq 1 - 1/2 - 1/5 \leq 1/4$, since $\eta < 1/10$ by assumption of the claim. Since

$$\begin{aligned}(1-\epsilon)\,(1-1/k-2\eta) - (1-\epsilon-4\eta)(1-1/k) &= (1-\epsilon)(1-1/k) - (1-\epsilon)2\eta \\ &\quad - (1-\epsilon)(1-1/k) + 4\eta(1-1/k) \\ &= 2\eta(-1+\epsilon+2(1-1/k)) \\ &\geq 2\eta\epsilon \geq 0 \quad \text{(since } k \geq 2\text{)},\end{aligned}$$

we in particular have

$$\mathbb{P}\left[|L^X| < (1-\epsilon-4\eta)\frac{1}{k}\,(1-1/k)\,|L^{X^*}|\right] \leq 2e^{-\epsilon^2|L^{X^*}|/(12k)}. \tag{22}$$

Now as long as $|L^{X^*}| > 48k\ln(1/\eta)/\epsilon^2$, we have that the rhs above is upper bounded by η^2. Since $|L^{X^*}| = W/(k|\mathbf{w}|)$ by Claim 20, this follows since $W \geq \frac{48k^2\ln(1/\eta)}{\epsilon^2} \cdot |\mathbf{w}|$ by assumption of the claim.

Finally, note that for every $\mathbf{w}$ we just showed that a single line L^X deviates from expectation with probability at most η^2. Since distinct lines do not overlap, an application of Chernoff bounds shows that for every $\mathbf{w}$ the probability that the number of lines $L^X \in \mathcal{L}^X(\mathbf{w})$ that deviate from expectation is at most $2\eta^2|\mathcal{L}^X(\mathbf{w})|$ with probability at least

$$1 - 2e^{-\Omega(\eta^2|\mathcal{L}^{X^*}|)} = 1 - 2e^{-\Omega(\eta^2 m^{4m}/(W/|\mathbf{w}|))}.$$

A union bound over at most 2^m vectors $\mathbf{w} \in \mathcal{F}$ and at most $(m^5)^{2^m}$ choices for the shifts $U(\mathbf{w}), \mathbf{w} \in \mathcal{F}$, yields failure probability at most $2 \cdot 2^m \cdot (m^5)^{2^m} \cdot e^{-\Omega(\eta^2 m^{4m}/(kW/|\mathbf{w}|))} < 1/100$ as long as m is sufficiently large as a function of η, k and $W/|\mathbf{w}|$. ■

Defining bipartite cliques induced by $L^Y \cup \pi^*(L^Y)$. We start with

Definition 22 (Typical lines) *For $\mathbf{w} \in \bigcup_{i=1}^k \mathcal{F}_i$ and $L^Y \in \mathcal{L}^Y(\mathbf{w})$ we say that L^Y and its pair $\pi_{\mathbf{w}}(L^Y)$ are* typical *if $|\pi_{\mathbf{w}}(L^Y)| \geq \frac{1}{k}(1-1/k)(1-4\eta-\epsilon)W/|\mathbf{w}|$.*

If L^Y is typical as per Definition 22, let $\tilde{L}^Y$ denote an arbitrary subset of L^Y of cardinality $(1-4\eta-\epsilon)|L^Y|$. Similarly, let $\tilde{L}^X$ denote an arbitrary subset of $\pi_{\mathbf{w}}(L^Y)$ of cardinality $(1-4\eta-\epsilon)(1-1/k)|L^Y|$. Our parameter setting will ensure that $|L^Y| = W/(k|\mathbf{w}|)$ is an integer multiple of $lcm(1/\eta, 1/\epsilon, k)$, so this is feasible. For convenience let $\tilde{L}^X = \tilde{L}^Y = \emptyset$ for lines that are not typical. We thus have $|\tilde{L}^X| = (1-1/k)|\tilde{L}^Y|$. Now let

$$\tilde{E}(L^Y) := \tilde{L}^Y \times \tilde{L}^X. \tag{23}$$

Note that for a typical line $L^Y \in \mathcal{L}^Y(\mathbf{w})$ the degree of a vertex $y \in L^Y$ in $\tilde{E}(L^Y)$ is either zero or $(1-4\eta-\epsilon)(1-1/k)|L^Y| = (1-4\eta-\epsilon)(1-1/k) \cdot \frac{W}{|\mathbf{w}|} =: (1-1/k)\gamma$, where we let

$$\gamma := (1-4\eta-\epsilon)\frac{W}{|\mathbf{w}|}, \tag{24}$$

and the degree of a vertex $x \in \pi(L^Y)$ is either zero or $(1-4\eta-\epsilon)|L^Y| = (1-4\eta-\epsilon)\frac{W}{|\mathbf{w}|} = \gamma$. Also note that all edges in the graph that we just defined are of the form (c, d), where

$$c = d + \lambda \cdot \mathbf{w}, |\lambda| \leq \frac{W}{|\mathbf{w}|}. \tag{25}$$

Defining the edges of $H_i^{\mathbf{w}}, \mathbf{w} \in \mathcal{F}_{i+1}$. Let

$$\tilde{E}_i^{\mathbf{w}} := \bigcup_{y \in T_i \setminus T_i^{\mathbf{w}}} \tilde{E}(\text{line}^Y(y, \mathbf{w})),$$

where $\tilde{E}(\text{line}^Y(y, \mathbf{w}))$ is defined in (23), T_i is defined in (12) and $T_i^{\mathbf{w}}$ is defined in (13). Note that since our vectors are only nearly orthogonal, for a $y \in T_i$ we do not necessarily have $\text{line}^Y(y, \mathbf{w}) \subseteq T_i$. We now let

$$E_i^{\mathbf{w}} := \tilde{E}_i^{\mathbf{w}} \cap (T_i \times S_i). \tag{26}$$

3.1.3 Induced property of subgraphs $H_i^{\mathbf{w}}$

We now show that the graphs $H_i^{\mathbf{w}}$ constructed above are induced for each i and $\mathbf{w} \in \mathcal{F}_i$. The argument is similar to [FLN$^+$02, GKK12].

Claim 23 *For every $\eta \in (0,1)$, integer $k \geq 2$, if $\epsilon \in (0,1)$ is smaller than η, then for every $i = 0, \ldots, k-1$, the edge set $E_i^{\mathbf{w}}$ (defined in* (26)*) is an induced union of subgraphs $H_i^{\mathbf{w}}$.*

Proof: Recall that $\mathcal{F}_{i+1}$ was chosen as a family of binary vectors of fixed weight with small intersections, namely for every $\mathbf{w}, \mathbf{w}' \in \mathcal{F}_{i+1}, \mathbf{w} \neq \mathbf{w}'$ one has

$$(\mathbf{w}, \mathbf{w}') \leq \epsilon |\mathbf{w}|. \tag{27}$$

Suppose that an edge $(c, d) \in E(H_i^{\mathbf{w}}), c \in X, d \in Y$ is induced by $H_i^{\mathbf{w}'}$ for $\mathbf{w}' \neq \mathbf{w}$. Since edges of $H_i^{\mathbf{w}'}$ connect red points in Y with respect to $\mathbf{w}'$ to blue points in X with respect to $\mathbf{w}'$ (see (19), (20) and the definition of edges in $H_i^{\mathbf{w}}$ in (23) and (26); see also (19) and (20)), it must be that $d \in R^Y(\mathbf{w}')$ and $c \in B^X(\mathbf{w}')$, so

$$|(c - d, \mathbf{w}')| \geq \eta \cdot W. \tag{28}$$

However, by (27) together with (25) one has

$$|(c - d, \mathbf{w}')| = |\lambda| \cdot (\mathbf{w}, \mathbf{w}') \leq \frac{W}{|\mathbf{w}|}(\mathbf{w}, \mathbf{w}') \leq \frac{W}{|\mathbf{w}|}\epsilon |\mathbf{w}| = \epsilon W < \eta W,$$

since $\epsilon < \eta$ by assumption of the claim. This yields a contradiction with (28), and hence $H_i^{\mathbf{w}}$ are induced. ■

3.1.4 Existence of a large matching in the host graph

We now show that with high probability over the choice of the random shifts $U(\mathbf{v}), \mathbf{v} \in \bigcup_{i=1}^k \mathcal{F}_i$, for any $i = 0, \ldots, k-1$ any collection $\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}, \mathbf{u}_s \in \mathcal{F}_s, s = 1, \ldots, i-1$ and $\mathbf{w} \in \mathcal{F}_{i+1}$ there exists a matching of $1 - O(k^3\epsilon/\eta)$ fraction of S_i to $T_i \setminus T_i^{\mathbf{w}}$. Formally we prove

Claim 24 *For every integer $k \geq 2$, sufficiently small $\eta \in (0,1)$ such that $1/\eta$ is an integer, $\epsilon \in (0, c \cdot \eta^2/k^6)$ for a sufficiently small constant $c > 0$ such that $1/\epsilon^{1/2}$ is an integer, if $\theta = \eta$ (see* (8)*) and W/w is an integer multiple of $k/(\epsilon \cdot \theta)$, the following conditions hold for sufficiently large m.*

There exists an event $\mathcal{E}_{balanced-degrees}$ that occurs with probability at least $99/100$ over the choice of random shifts $U(\mathbf{v}), \mathbf{v} \in \bigcup_{j=1}^k \mathcal{F}_j$, for every $i = 0, \ldots, k-1$, every collection $\mathbf{u}_1, \ldots, \mathbf{u}_s \in \mathcal{F}_s, s = 1, \ldots, i$ and every $\mathbf{w} \in \mathcal{F}_{i+1}$ such that conditioned on $\mathcal{E}_{balanced-degrees}$ and the event $\mathcal{E}$ from Lemma 19 there exists a matching of $1 - O(k^3\epsilon^{1/2}/\eta)$ fraction of S_i to $T_i \setminus T_i^{\mathbf{w}}$.

Proof: We will do this by exhibiting a fractional matching of appropriate size. Recall that a fractional matching is an assignment of non-negative weights z_e to edges e of the graph such that for every vertex v of the graph one has $\sum_{e \in \delta(v)} z_e \leq 1$. We now exhibit a fractional matching in the graph in three steps.

First, for every typical line $L^Y \in \mathcal{L}^Y(\mathbf{w})$ that touches $T_i \setminus T_i^{\mathbf{w}}$ we assign weights to every edge of $\tilde{E}(L^Y)$ in such a way that every vertex in L^Y that has nonzero degree in $\tilde{E}(L^Y)$ receives $1 - 1/k$ fractional mass, and every vertex in $\pi(L^Y)$ that has a nonzero degree receives mass 1. Then we assign fractional mass uniformly to edges incident on vertices in $S_i \setminus S_i^{\mathbf{w}}$ to ensure that these vertices contribute the missing $1/k$ fraction of mass to vertices in L^Y, up to a small error term that is independent of k, the number of rounds in the game, and can be made arbitrarily small by choosing the maximum dot product ϵ between vectors in $\bigcup_{j=1}^k \mathcal{F}_j$ small, and making the 'buffer' between red and blue vertices appropriately small (this mass is assigned to edges in lines $L^Y \in \mathcal{L}^Y(\mathbf{v})$ for $\mathbf{v} \in \mathcal{F}_{i+1} \setminus \{\mathbf{w}\}$). This ensures that the matching supported by the lines that touch $T_i \setminus T_i^{\mathbf{w}}$ is about the size of S_i. The only problem is that this matching uses edges outside of $T_i \setminus T_i^{\mathbf{w}}$ and S_i. We then show that pruning to edges contained in $(T_i \setminus T_i^{\mathbf{w}}) \times S_i$ only affects matching size by a small error term, completing the proof.

Step 1: weights on edges of $H_i^{\mathbf{w}}$. Recalling that for a typical line $L^Y \in \mathcal{L}^Y(\mathbf{w})$ the degree of every vertex in L^Y is either zero or $(1 - 1/k)\gamma$ (where γ is defined in (24)), we put weight $1/\gamma$ on every edge of $\tilde{E}(L^Y)$. This way every vertex of nonzero degree in L^Y gets fractional mass $1 - 1/k$, and every vertex of nonzero degree in $\pi(L^Y)$ gets fractional mass 1.

Step 2: weights on edges of $H_i^{\mathbf{v}}$ for $\mathbf{v} \neq \mathbf{w}$. We start by showing that for a fixed $\mathbf{v}$ and for every $y \in Y$ one has that $\mathbb{P}_{U(\mathbf{v})}[y \in R^Y(\mathbf{v})]$ is very close to $\frac{1}{k}$. Indeed, recall that $U(\mathbf{v})$ is uniformly random over the set

$$\{0, \ldots, k/\theta - 1\} \cdot W \cdot (\theta/k),$$

where $\theta \in (0, 1)$ is a parameter that by assumptions of the lemma is equal to η (see (8)). Using the definition of $R^Y(\mathbf{v})$ (see (9)) we can now bound

$$\mathbb{P}_{U(\mathbf{v})}[y \in R^Y(\mathbf{v})] = \mathbb{P}_{U(\mathbf{v})}[((y, \mathbf{v}) + U(\mathbf{v})) \mod W \in [0, 1/k) \cdot W].$$

Writing $(y, \mathbf{v}) = \left\lfloor \frac{(y,\mathbf{v})}{W \cdot (\theta/k)} \right\rfloor \cdot W \cdot (\theta/k) + ((y, \mathbf{v}) \bmod (W \cdot (\theta/k)))$ and recalling that $U(\mathbf{v})$ is uniformly random in $\{0, \ldots, k/\theta - 1\} \cdot W \cdot (\theta/k)$ by definition as well as that $1/\theta$ is an integer, we get that

$$\mathbb{P}_{U(\mathbf{v})}[((y, \mathbf{v}) + U(\mathbf{v})) \mod W \in [0, 1/k) \cdot W] = (1/\theta)/(k/\theta) = 1/k,$$

as required. A similar argument shows that for every $x \in S_i$ one has $\mathbb{P}_{U(\mathbf{v})}[x \in B^{X^*}(\mathbf{v})] = 1 - 1/k - 2\eta$. Indeed, this is because

$$\mathbb{P}_{U(\mathbf{v})}[((x, \mathbf{v}) + U(\mathbf{v})) \mod W \in [1/k + \eta, 1 - \eta) \cdot W] = (k/\theta)(1 - 1/k - 2\eta)/(k/\theta) = 1 - 1/k - 2\eta,$$

where we used the assumption that $\theta = \eta$.

Next note that each vertex $y \in R^Y(\mathbf{w}) \setminus \text{Bad}$ has degree $(1 - 1/k)\gamma$ or 0 in $H_i^{\mathbf{v}}$, and for every $\mathbf{v}$ the fraction of vertices that have degree 0 in $H_i^{\mathbf{v}}$ is at most $2\eta^2|Y|$ by Claim 21. Furthermore, since the random shifts $U(\mathbf{v})$ are independent for distinct $\mathbf{v}$, we obtain using Chernoff bounds (Theorem 5) for $\delta \in (0, 1)$ that for every $\mathbf{w} \in \mathcal{F}_{i+1}$ and every $y \in T_i \setminus T_i^{\mathbf{w}}$

$$\mathbb{P}_{\{U(\mathbf{v})\}_{\mathbf{v} \in \mathcal{F}_{i+1}, \mathbf{v} \neq \mathbf{w}}} \left[\sum_{\mathbf{v} \in \mathcal{F}_{i+1}, \mathbf{v} \neq \mathbf{w}} \mathbf{I}[y \in R^Y(\mathbf{v})] \notin (1 \pm \delta) d/k \right] \leq 2e^{-\delta^2 d/(3k)}.$$

Similarly we have for every $\mathbf{w} \in \mathcal{F}_{i+1}$ and $x \in S_i \setminus S_i^{\mathbf{w}}$ and $\delta \geq 4\eta$

$$\mathbb{P}_{\{U(\mathbf{v})\}_{\mathbf{v}\in\mathcal{F}_{i+1},\mathbf{v}\neq\mathbf{w}}}\left[\sum_{\mathbf{v}\in\mathcal{F}_{i+1},\mathbf{v}\neq\mathbf{w}} \mathbf{I}[x \in B^X(\mathbf{v})] \notin (1\pm\delta)d(1-1/k)\right] \leq 2e^{-\delta^2 d/24}.$$

Let $\mathcal{E}_{balanced-degrees}$ denote the event that for every every collection $\mathbf{u}_1,\ldots,\mathbf{u}_{i-1},\mathbf{u}_s \in \mathcal{F}_s, s=1,\ldots,i-1$, every $\mathbf{w}\in\mathcal{F}_{i+1}$, every $y \in T_i$ one has $\sum_{\mathbf{v}\in\mathcal{F}_{i+1},\mathbf{v}\neq\mathbf{w}} \mathbf{I}[y \in R^Y(\mathbf{v})] \in (1\pm\delta)d/k$ (i.e. y is a red vertex with respect to about the expected number of vectors $\mathbf{v}$) and $\sum_{\mathbf{v}\in\mathcal{F}_{i+1},\mathbf{v}\neq\mathbf{w}} \mathbf{I}[x \in B^X(\mathbf{v})] \notin (1\pm\delta)d(1-1/k)$ (i.e. x is a blue vertex with respect to about the expected number of vectors $\mathbf{v}$). Since there are only $O(m^{4m})$ vertices in Y and X, and $|\bigcup_i \mathcal{F}_i| \leq 2^m$, and $d = 2^{\Omega(m)}$, for any constant k,η,δ and sufficiently large m a union bound shows that $\mathcal{E}_{balanced-degrees}$ occurs with probability at least $99/100$.

The assignment of fractional weights on edges incident to vertices in $S_i \setminus S_i^{\mathbf{w}}$ is as follows: we put weight $\frac{1}{(1-1/k)\gamma\cdot(1+\delta)d}$ on each edge of $\tilde{E}(L^Y)$ for $L^Y \in \mathcal{L}^Y(\mathbf{v})$ that is incident on $y \in T_i \setminus T_i^{\mathbf{w}}$. We now verify feasibility of this solution in the presence of weights assigned in step 1, and then compute the size of the matching.

To verify feasibility, note that, conditioned on $\mathcal{E}_{balanced-degrees}$, the contribution of this assignment to any vertex in $S^i \setminus S^i(\mathbf{w})$ is at most

$$\frac{1}{(1-1/k)\gamma\cdot(1+\delta)d}\cdot(1+\delta)d(1-1/k)\cdot\gamma = 1,$$

where we used the fact that the degree of a vertex in $S_i \setminus S_i^{\mathbf{w}}$ in a subgraph induced by a typical line is at most γ. Contribution to any vertex in $T_i \setminus T_i^{\mathbf{w}}$ is at most

$$\frac{1}{(1-1/k)\gamma\cdot(1+\delta)d}\cdot(1+\delta)d/k\cdot(1-1/k)\gamma = 1/k,$$

where we used the fact that the degree of a vertex in $T_i \setminus T_i^{\mathbf{w}}$ in a subgraph induced by a typical line is at most $(1-1/k)\gamma$. Thus, the total mass assigned to vertices in $T_i \setminus T_i^{\mathbf{w}}$ as well as $S_i \setminus S_i^{\mathbf{w}}$ is upper bounded by 1.

To lower bound the value of the fractional solution, first note that by Claim 21 with high probability over the choice of X for every $\mathbf{w}$ at most $2\eta^2|Y|$ points belong to atypical lines (see Definition 22), which corresponds to a loss of at most $2\eta^2|Y|$ in matching size. Now recall that a line is called typical (see Definition 22) if $|\pi_{\mathbf{w}}(L^Y)| \geq (1-1/k-4\eta-\epsilon)W/|\mathbf{w}|$. Thus, at most a $4\eta+\epsilon$ fraction of mass assigned is lost due to this. Since this applies to every $\mathbf{v}\in\mathcal{F}_{i+1}, \mathbf{v}\neq\mathbf{w}$, as well, we get that the constructed fractional matching is feasible, and its size is at least $(1-O(\delta+\epsilon))|S_i| - O(\eta^2)|Y| = (1-O(\eta k))|S_i|$, where we set $\delta = 4\eta$ and used the fact that conditioned on the event $\mathcal{E}$ from Lemma 19 one has $|S_i| = \Omega(1/k)|Y|$.

Step 3: bounding effect of truncation of $\tilde{E}_i^{\mathbf{w}}$ to $E_i^{\mathbf{w}}$ In steps 1 and 2 we showed that the mass we assigned to $\tilde{E}_i^{\mathbf{w}}$ corresponds to a feasible matching of size at least $(1-O(\eta k))|S_i|$. We now show that truncating $\tilde{E}_i^{\mathbf{w}}$ to $E_i^{\mathbf{w}}$ (see (26)) does not lead to a significant loss in matching size. Recall that by (12)

$$T_i = \{y \in Y : ((y,\mathbf{u}_j)+U(\mathbf{u}_j)) \mod W \in [1/k,1)\cdot W, \text{ for all } j\in[1:i]\}$$
$$S_i = \{x \in X : ((x,\mathbf{u}_j)+U(\mathbf{u}_j)) \mod W \in [1/k,1)\cdot W, \text{ for all } j=[1:i]\}$$

For every $y \in T_i$ one has by (12) that $((y,\mathbf{u}_j)+U(\mathbf{u}_j)) \mod W \in [1/k,1)\cdot W$ for all $j=1,\ldots,i$, and hence for every $\lambda \in (0, W/|\mathbf{w}|]$ and every $j=1,\ldots,i$

$$|(y+\lambda\mathbf{w},\mathbf{u}_j) - (y,\mathbf{u}_j)| \leq \lambda(\mathbf{w},\mathbf{u}_j) \leq \epsilon\lambda|\mathbf{w}| \leq \epsilon W$$

by choice of the family $\mathcal{F}_1, \ldots, \mathcal{F}_k$. We thus get that $y + \lambda \mathbf{w}$ belongs to the set

$$\hat{T}_i := \{y \in Y : ((y, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in ([0, \epsilon) \cup [1/k - \epsilon, 1)) \cdot W, \text{ for all } j \in [1 : i]\},$$

i.e. the result of relaxing the constraints that define T_i by a ϵ in every direction. At the same time

$$\begin{aligned}\hat{T}_i \setminus T_i &\subseteq \bigcup_{j=1}^{i} \{y \in Y : ((y, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in ([0, \epsilon) \cup [1/k - \epsilon, 1/k)) \cdot W\} \\ &\subseteq \bigcup_{j=1}^{i} \{y \in Y : ((y, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in ([0, \epsilon^{1/2}) \cup [1/k - \epsilon^{1/2}, 1/k)) \cdot W\}\end{aligned}$$

and thus

$$\begin{aligned}|\hat{T}_i \setminus T_i| &\leq \sum_{j=1}^{i} |\{y \in Y : ((y, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in ([0, \epsilon^{1/2}) \cup [1/k - \epsilon^{1/2}, 1/k)) \cdot W\}| \\ &\leq 2k(6(k/\theta) \cdot \epsilon^{1/2} + 4/m)|Y| \qquad \text{(by Lemma 18)}\end{aligned}$$

To obtain the last bound, we applied Lemma 18 for each $j = 1, \ldots, i$ with $\mathcal{U} = \{\mathbf{u}_j\}$, $\delta' = \epsilon$ and $L = (k/\theta) \cdot \epsilon^{-1/2}$. Now recalling that by (26) one has $E_i^{\mathbf{w}} := \tilde{E}_i^{\mathbf{w}} \cap (T_i \times S_i)$ and that S_i is obtained by intersecting T_i with X, we get that the edges pruned from $\tilde{E}_i^{\mathbf{w}}$ by restricting to $T_i \times S_i$ as above are incident on as set of vertices of size at most

$$2|\hat{T}_i \setminus T_i| \leq 4k(6(k/\theta) \cdot \epsilon^{1/2} + 4/m)|Y|. \tag{29}$$

Since every vertex received at most 1 unit of fractional mass, the size of the matching supported by $E_i^{\mathbf{w}}$ is thus at least the size of the matching supported by $\tilde{E}_i^{\mathbf{w}}$ minus $2k \cdot (6(k/\theta) \cdot \epsilon^{1/2} + 4/m)|Y| = O(k^2\epsilon^{1/2}/\eta)|Y| = O(k^3\epsilon^{1/2}/\theta)|S_i|$ giving the result. In the last transition we used the fact that $|S_i| = \Omega(|Y|/k)$ when $\epsilon < c \cdot \theta^2/k^6$ for a sufficiently small constant $c > 0$ (by Lemma 19), as well as the assumption that $\theta = \eta$.

■

3.1.5 Existence of a sparse directed cut

Define

$$Z := \{y \in Y : ((y, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in ([1/k - \epsilon, 1/k) \cup [0, \epsilon)) \cdot W \text{ for some } j \in [1 : k]\}. \tag{30}$$

We prove

Claim 25 *For every integer $k \geq 2$, $\epsilon \in (0, 1)$ such that $1/\epsilon^{1/2}$ is an integer, if W/w is an integer multiple of $k/(\epsilon \cdot \theta)$, the following conditions hold for sufficiently large m.*

For every $i = 0, 1, \ldots, k-1$ the subgraph H^* induced by $(T_i \setminus (T_k \cup Z)) \cup S_i^*$ only contains the edges of $E_i^{\mathbf{u}_{i+1}}$. In addition, one has $|Z| \leq 2k^2(6\epsilon^{1/2}/\theta + 4/m)|Y|$.

Proof: Recall that the sets S_i^* are defined in (14). First note that if an edge $(c, d), c \in P, d \in Q$ belongs to H^*, then $c \in S_i^*$ and $d \in T^i$, so (c, d) necessarily belongs to some graph $H_i^{\mathbf{w}}$, where $\mathbf{w} \in \mathcal{F}_{i+1}$. Then we have by (25) that

$$d - c = \lambda \cdot \mathbf{w}, \text{ where } |\lambda| \leq W/|\mathbf{w}|.$$

Thus, we have for all $j = 1, \ldots, k$ using the orthogonality condition (27)

$$|(c - d, \mathbf{u}_j)| \leq \frac{W}{|\mathbf{w}|} |(\mathbf{w}, \mathbf{u}_j)| \leq \epsilon W. \tag{31}$$

Now recall that $c \in S_i^*$ by assumption, so by (12) and (14)

$$(c, \mathbf{u}_j) + U(\mathbf{u}_j) \mod W \in [1/k, 1) \cdot W, \forall j = 1, \ldots, k.$$

Thus, by (31) one has

$$(d, \mathbf{u}_j) + U(\mathbf{u}_j) \mod W \in ([1/k - \epsilon, 1] \cup [0, \epsilon)) \cdot W, \text{ for all } j = 1, \ldots, k,$$

i.e. $d \in Z \cup T^k$. Thus, the subgraph H^* contains only edges of $E_i^{\mathbf{u}_{i+1}}$ for every $i = 0, \ldots, k-1$, as required.

It remains to bound the size of Z. We first note that

$$\begin{aligned} Z &= \{y \in Y : ((y, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in ([1/k - \epsilon, 1/k) \cup [0, \epsilon)) \cdot W \text{ for some } j \in [1:k]\} \\ &\subseteq \bigcup_{j=1}^{k} \{y \in Y : ((y, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in ([1/k - \epsilon, 1/k) \cup [0, \epsilon)) \cdot W\} \\ &\subseteq \bigcup_{j=1}^{k} \{y \in Y : ((y, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in ([1/k - \epsilon^{1/2}, 1/k) \cup [0, \epsilon^{1/2})) \cdot W\}. \end{aligned} \tag{32}$$

For each $j = 1, \ldots, k$ we now use Lemma 18 with $\mathcal{U} = \{\mathbf{u}_j\}$, $a_{\mathbf{u}_j} = \frac{1}{k} - \epsilon, b_{\mathbf{u}_j} = \frac{1}{k}$, and then again with $a_{\mathbf{u}_j} = 0, b_{\mathbf{u}_j} = \epsilon$. In both cases we set $L = (k/\theta) \cdot 1/\epsilon^{1/2}$. We thus get

$$\begin{aligned} &|\{y \in Y : ((y, \mathbf{u}_j) + U(\mathbf{u}_j)) \mod W \in ([1/k - \epsilon^{1/2}, 1/k) \cup [0, \epsilon^{1/2})) \cdot W\}| \\ &\leq 2(6L\epsilon + 4/m) \cdot |Y| \\ &\leq 2(6k\epsilon^{1/2}/\theta + 4/m) \cdot |Y|. \end{aligned}$$

Using this together with (32) yields $|Z| \leq 2k^2(6\epsilon^{1/2} + 4/m)|Y|$, as required.

■

3.2 Distribution over inputs

We now formally define our hard input distribution. The input graph G' is generated as follows. First sample $X \subseteq \{0,1\}^m$ as in Definition 17, then for every $\mathbf{w} \in \bigcup_{j=1}^{k} \mathcal{F}_j$ sample the shift $U(\mathbf{w})$ independently as per (8). Finally, sample $\mathbf{u}_s \in \mathcal{F}_s, s = 1, \ldots, k$ independently and uniformly at random, and let $G := G(\mathbf{u}_1, \ldots, \mathbf{u}_k)$ denote the host graph as constructed in Section 3.1. For every $\mathbf{v} \in \mathcal{F}_{i+1}$ and $y \in T_i \setminus T_i^{\mathbf{v}}$ let $b_y^{\mathbf{v}} \in \{0,1\}$ denote a Bernoulli random variable with expectation $1 - \xi$ for a small $\xi \in (0,1)$ that we will set later. The variables $b_y^{\mathbf{v}}$ are independent conditioned on

$$\sum_{y \in T_i \setminus T_i^{\mathbf{v}}} b_y^{\mathbf{v}} = \lceil (1-\xi)|T_i \setminus T_i^{\mathbf{v}}| \rceil$$

for every $\mathbf{v} \in \mathcal{F}_{i+1}$. In other words, $b^{\mathbf{v}}$ encodes a uniformly random subset of $T_i \setminus T_i^{\mathbf{v}}$ of size $\lceil (1-\xi)T_i \setminus T_i^{\mathbf{v}} \rceil$. For every $i = 0, \ldots, k-1$ let $B_i := \{b_y^{\mathbf{v}}\}_{\mathbf{v} \in \mathcal{F}_{i+1}, y \in T_i \setminus T_i^{\mathbf{v}}}$.

Definition 26 (Subsampling of the host graph) *For $i = 0, \ldots, k-1$ let the graph $G'(\mathbf{u}_{1:i}; B_{0:i})$ be formed by including, for every $\mathbf{v} \in \mathcal{F}_{i+1}$ and $y \in T_i \setminus T_i^{\mathbf{v}}$ all edges incident on y in $H_i^{\mathbf{v}}$ if $b_y^{\mathbf{v}} = 1$ and none of these edges otherwise. For $i = k$, let $G'(\mathbf{u}_{1:k}; B_{0:k-1})$ contain all edges incident on S_k. Let*

$$G' := \left(\bigcup_{i=0}^{k-1} G'(\mathbf{u}_{1:i}; B_{0:i}) \right) \cup G'(\mathbf{u}_{1:k}; B_{0:k-1}).$$

The stream consists of $k+1$ *phases*: for each $i = 0, \ldots, k$ the vertices and edges of $G'(\mathbf{u}_{1:i}, B_{0:i})$ incident on S_i arrive in phase i in an arbitrary order.

We start with the following claim

Claim 27 *For every $\xi \in (0,1)$, every integer $k \geq 2$, sufficiently small $\eta \in (0,1)$ such that $1/\eta$ is an integer, $\epsilon \in (0, c \cdot \eta^2/k^6)$ for a sufficiently small constant $c > 0$ such that $1/\epsilon^{1/2}$ is an integer, if $\theta = \eta$ (see (8)) and W/w is an integer multiple of $k/(\epsilon \cdot \theta)$, the following conditions hold for sufficiently large m.*

There exists an event $\mathcal{E}_{large-matching}$ that occurs with probability at least $97/100$ over the choice of random shifts $U(\mathbf{v}), \mathbf{v} \in \bigcup_{j=1}^{k} \mathcal{F}_j$ and choice of $\bigcup_{i=0}^{k-1} \{b_y^{\mathbf{v}}\}_{\mathbf{v} \in \mathcal{F}_{i+1}, y \in T_i \setminus T_i^{\mathbf{v}}}$, the graph G' contains a matching of size at least $(1 - O(\xi + k^4 \epsilon^{1/2}/\eta))|Y|$.

Proof: By Claim 24 there exists an event $\mathcal{E}_{balanced-degrees}$ that depends only on the choice of X and the shifts $U(\mathbf{w}), \mathbf{w} \in \bigcup_{j=1}^{k} \mathcal{F}_j$ such that conditioned on $\mathcal{E}_{balanced-degrees}$ and event $\mathcal{E}$ from Lemma 19 by Claim 24 for every $i = 0, \ldots, k-1$ and $\mathbf{u}_1, \ldots, \mathbf{u}_s \in \mathcal{F}_s, s = 1, \ldots, i$ and $\mathbf{w} \in \mathcal{F}_{i+1}$ there exists a matching M_i of $1 - O(k^3 \epsilon^{1/2}/\eta)$ fraction of S_i to $T_i \setminus T_i^{\mathbf{u}_{i+1}}$. Furthermore, the set S_k can be perfectly matched to T_k by definition. Let M be a union of these matchings. Note that conditioned on $\mathcal{E}_{balanced-degrees}$ the matching M satisfies

$$\begin{aligned}
|M| &\geq (1 - O(\xi + k^3 \epsilon^{1/2}/\eta))|S| + |T_k| \\
&\geq (1 - O(\xi + k^3 \epsilon^{1/2}/\eta)) \sum_{i=0}^{k-1} |S_i| + |T_k| \\
&\geq (1 - O(\xi + k^3 \epsilon^{1/2}/\eta)) \sum_{i=0}^{k-1} (1 - 1/k)^i |Y|/k + |T_k| - O(k^4 \epsilon/\theta)|Y| \\
&\geq (1 - O(\xi + k^3 \epsilon^{1/2}/\eta))(1 - (1 - 1/k)^k)|Y| + (1 - 1/k)^k |Y| - O(k^4 \epsilon/\theta) \\
&\geq (1 - O(\xi + k^4 \epsilon^{1/2}/\eta))|Y|,
\end{aligned}$$

where we used Lemma 19, **(2)**, in the third transition and Lemma 19, **(1)**, in the forth transition.

Now recall that G' contains every edge of M with probability at least $1 - \xi$, and these events are negatively associated for different edges, since for every $(y, x) \in M \cap H^{\mathbf{u}_{i+1}}$ one has $(y, x) \in G'$ if and only if $b_y^{\mathbf{u}_{i+1}} = 1$, and $b_y^{\mathbf{u}_{i+1}}$ are negatively associated for different y by construction. We thus get by an application of Chernoff bounds that

$$\mathbb{P}[|M \cap G'| < (1 - 2\xi)|M|] < e^{-\Omega(\xi^2 |M|)} < e^{-\Omega(\xi^2 m^{4m})}.$$

We now define the event $\mathcal{E}_{large-matching}$ to be the intersection of $\mathcal{E}_{balanced-degrees}$ with the success events for $i = 0, \ldots, k-1$ above, getting that $\mathbb{P}[\mathcal{E}_{large-matching}] \geq 1 - 97/100$ by a union bound.

■

3.3 Bounding performance of a small space algorithm

By Yao's minimax principle it is sufficient to upper bound the performance of a deterministic small space algorithm that succeeds with probability at least $1/2$. To do that, we bound the size of the matching that a small space algorithm can output at the end of the stream. Let M_{ALG} denote the matching that the algorithm outputs. We first upper bound the approximation ratio that the algorithm obtains in terms of the number of edges in $E(H_i^{\mathbf{u}_{i+1}}) \cap M_{ALG}$, for $i = 0, 1, \ldots, k-1$.

Lemma 28 *For every integer $k \geq 2$, $\epsilon \in (0,1)$ such that $1/\epsilon^{1/2}$ is an integer, if W/w is an integer multiple of $k/(\epsilon \cdot \theta)$ (see (8)), the following conditions hold for sufficiently large m.*

If the graph G' is generated as per Definition 26, and M_{ALG} is any matching in G', then $|M_{ALG}| \leq (1 - 1/k)^k |Y| + \sum_{i=0}^{k-1} |E(H_i^{\mathbf{u}_{i+1}}) \cap M_{ALG}| + O(k^3 \epsilon^{1/2}/\theta) \cdot |Y|$.

Proof: Consider the cut (A, B), where $A = (T \setminus (T_k \cup Z)) \cup \bigcup_{i=0}^{k-1}(S_i \setminus S_i^*)$ and $B = T_k \cup S_k \cup \bigcup_{i=0}^{k-1} S_i^* \cup Z$. Recall that the sets T_i, S_i are defined in (12), S_i^j is defined in (14) and Z is defined in (30).

By the maxflow/mincut theorem, the size of the matching output by the algorithm is bounded by $|A \cap S| + |B \cap T| + |((A \cap T) \times (B \cap S)) \cap M_{ALG}|$.

By Claim 25 the subgraph $M_{ALG} \cap (T_i \times S_i)$ induced by $(T_i \setminus (T_k \cup Z)) \cup S_i^*$ only contains the edges of $H_i^{\mathbf{u}_{i+1}}$ for every $i = 0, \ldots, k-1$. Thus,

$$|((A \cap T) \times (B \cap S)) \cap M_{ALG}| \leq \sum_{i=0}^{k-1} |E(H_i^{\mathbf{u}_{i+1}}) \cap M_{ALG}|$$

and we get

$$\begin{aligned} |M_{ALG}| &\leq \left|\bigcup_{i=0}^{k-1}(S_i \setminus S_i^*)\right| + |T_k| + |Z| + \sum_{i=0}^{k-1} |E(H_i^{\mathbf{u}_{i+1}}) \cap M_{ALG}| \\ &= \left|\bigcup_{i=0}^{k-1} S_i\right| + |T_k| - \left|\bigcup_{i=0}^{k-1} S_i^*\right| + |Z| + \sum_{i=0}^{k-1} |E(H_i^{\mathbf{u}_{i+1}}) \cap M_{ALG}| \end{aligned} \tag{33}$$

Furthermore, again by Claim 25 one has $|Z| \leq 2k^2(6\epsilon^{1/2}/\theta + 4/m)|Y|$. By Lemma 19, **(1)** one has $|T_i| = (1-1/k)^i|Y| + \Delta_i$, $|\Delta_i| = O(k^3\epsilon/\theta) \cdot |Y|$ for every $i \in \{0, 1, 2, \ldots, k\}$. By Lemma 19, **(2)** one has $|S_i| = \frac{1}{k}((1-1/k)^i|Y| + \Delta_i$, $|\Delta_i| = O(k^3\epsilon/\theta) \cdot |Y|$ for every $i \in \{0, 1, 2, \ldots, k\}$ and by Lemma 19, **(3)** one has $|S_i^*| = \frac{1}{k}(1-1/k)^k|Y| + \Delta_i$, $|\Delta_i| = O(k^3\epsilon/\theta) \cdot |Y|$ for every $i \in \{0, 1, 2, \ldots, k\}$. Substituting these bounds into (33), we get

$$\begin{aligned} |M_{ALG}| \leq &\sum_{i=0}^{k-1} \frac{1}{k}(1-1/k)^i|Y| - \sum_{i=0}^{k-1} \frac{1}{k}(1-1/k)^k|Y| \\ &+ (1-1/k)^k \cdot |Y| + \sum_{i=0}^{k-1} |E(H_i^{\mathbf{u}_{i+1}}) \cap M_{ALG}| + O(k^3\epsilon^{1/2}/\theta) \cdot |Y|. \end{aligned}$$

We now use the fact that $\sum_{i=0}^{k-1} \frac{1}{k}(1-1/k)^i|Y| = (1-(1-1/k)^k)|Y|$ in the upper bound above to get

$$|M_{ALG}| \leq (1-(1-1/k)^k)|Y| + \sum_{i=0}^{k-1} |E(H_i^{\mathbf{u}_{i+1}}) \cap M_{ALG}| + O(k^3\epsilon^{1/2}/\theta) \cdot |Y|,$$

as required. ∎

We will use

Lemma 29 (Data Processing Inequality) *For any random variables (X, Y, Z) such that $X \to Y \to Z$ forms a Markov chain, we have $I(X; Z) \leq I(X; Y)$.*

Lemma 30 *For every integer $k \geq 2$, if $\epsilon < \eta$, the following conditions hold for sufficiently large m.*

Let M_{ALG} denote the subset of edges of G' output by a space s streaming algorithm after a single pass over the edges of G' presented in the order defined above. If

$$\left|M_{ALG} \cap \bigcup_{i=0}^{k-1} E(H_i^{\mathbf{u}_{i+1}})\right| > cn$$

with probability more than $1/2$ over the randomness used to generate the graph G', then $s = \Omega_{c,k}(nd)$.

Proof: Recall that we define M_{ALG} to be the empty set if the matching output by ALG contains edges that are not in G, and that

$$\mathbb{P}\left[\left|M_{ALG} \cap \bigcup_{i=0}^{k-1} E(H_i^{\mathbf{u}_{i+1}})\right| > cn\right] > 1/2, \tag{34}$$

where the probability is over the choice of $\mathbf{u}_{1:k}$ (the vectors defining the host graph G) and $\mathbf{B}_{0:k-1}$ (the random variables used to subsample the host graph G to generate G'), subsampling X and the shifts $\{U(\mathbf{w})\}$. Define, for $\mathbf{v} \in \mathcal{F}_{i+1}$,

$$M_{ALG}^{\mathbf{v}} = M_{ALG} \cap E(H_i^{\mathbf{v}}).$$

Note that $M_{ALG}^{\mathbf{v}}$ depends on i, since $\mathbf{v}$ uniquely determines i (since it belongs to $\mathcal{F}_{i+1}$ and no other $\mathcal{F}_j, j = 0, 1, \ldots, k-1$). Also define

$$\mathcal{E}_{many-edges}(i) := \left\{|M_{ALG}^{\mathbf{v}}| > cn/(6k)\right\}.$$

We have, using (34), that there exists an index $0 \le i \le k-1$ such that

$$\mathbb{P}_{\mathbf{u}_{1:k}, \mathbf{B}_{0:k-1}, \mathbf{X}, \{\mathbf{U}(\mathbf{w})\}}[\mathcal{E}_{many-edges}(i)] \ge 1/(6k).$$

Fix one such index i in what follows. We have by an averaging argument applied to (34) (using Claim 23 to conclude that the edge sets of $H_i^{\mathbf{w}}$ are disjoint for distinct $\mathbf{w} \in \mathcal{F}_{i+1}$) that there exists a fixing $F = (u_{1:i}, B_{0:i-1}, X, \{U(\mathbf{w})\})$ of $\mathbf{u}_{1:i}, \mathbf{B}_{0:i-1}, \mathbf{X}, \{\mathbf{U}(\mathbf{w})\}$ such that

$$\mathbb{P}_{\mathbf{u}_{i+1:k-1}, \mathbf{B}_{i:k-1}}[\mathcal{E}_{many-edges}(t)|F] \ge 1/(6k).$$

For every $\mathbf{v} \in \mathcal{F}_{i+1}$ define

$$\mathcal{E}_{many-edges}(i, \mathbf{v}) := \mathcal{E}_{many-edges}(i) \wedge \{\mathbf{u}_{i+1} = \mathbf{v}\}.$$

We have

$$\begin{aligned} 1/(6k) \le & \mathbb{P}_{\mathbf{u}_{i+1:k-1}, \mathbf{B}_{i:k-1}}[\mathcal{E}_{many-edges}(i)|F] \le \mathbb{E}_{\mathbf{u}_{i+1}}[\mathbb{P}_{\mathbf{u}_{i+2:k}, \mathbf{B}_{i:k-1}}[\mathcal{E}_{many-edges}(i)|F]] \\ &= \frac{1}{|\mathcal{F}_{i+1}|} \sum_{\mathbf{v} \in \mathcal{F}_{i+1}} \mathbb{P}_{\mathbf{u}_{i+2:k}, \mathbf{B}_{i:k-1}}[\mathcal{E}_{many-edges}(i, \mathbf{v})|F]. \end{aligned} \tag{35}$$

Let Π denote the state of the algorithm after processing

$$\left(\bigcup_{j=0}^{i-1} G'(u_{1:j}; B_{0:j})\right) \cup G'(u_{1:i}; B_{0:i-1}, \mathbf{B}_i),$$

where $G'(u_{1:j}; B_{0:j})$ is given by Definition 26. We now lower bound $I(\Pi; \mathbf{B}_i)$ (which then gives a lower bound on the entropy of Π, and therefore on the space s). First note that

$$\mathbf{B}_i = \{b_i^{\mathbf{v}}\}_{\mathbf{v} \in \mathcal{F}_{i+1}} \to \Pi \to M_{ALG}$$

forms a Markov chain, and thus by the data processing inequality (Lemma 29) we have

$$I(\Pi; \mathbf{B}_i) \ge I(M_{ALG}; \mathbf{B}_i). \tag{36}$$

It thus suffices to lower bound $I(M_{ALG}; \mathbf{B}_i) = H(\mathbf{B}_i) - H(\mathbf{B}_i|M_{ALG})$. We upper bound the second term. First let E denote the indicator random variable of $\mathcal{E}_{many-edges}(i)$ conditioned on $\{\mathbf{u}_{1:i} = u_{1:i}$ and $\mathbf{B}_{0:i-1} =$

$B_{0:i-1}\}$ (note that by choice of the index i we have $\mathbb{E}_{\mathbf{u}_{i+1:k-1},\mathbf{B}_{i:k-1}}[E] \geq 1/(6k)$. Further, for every $\mathbf{v} \in \mathcal{F}_{i+1}$ let $E^{\mathbf{v}}$ denote the indicator random variable of $\mathcal{E}_{many-edges}(i,\mathbf{v})$ conditioned on $\{\mathbf{u}_{1:i} = u_{1:i}, \mathbf{B}_{0:i-1} = B_{0:i-1}\}$. Note that by (35) we have

$$\mathbb{E}_{\mathbf{v}\sim UNIF(\mathcal{F}_{i+1})}[E^{\mathbf{v}}] \geq 1/(6k). \tag{37}$$

$$\begin{aligned}
H(\mathbf{B}_i|\Pi) &\leq \sum_{\mathbf{v}\in\mathcal{F}_{i+1}} H(b^{\mathbf{v}}|\Pi) \qquad \text{(by subadditivity of entropy)} \\
&\leq \sum_{\mathbf{v}\in\mathcal{F}_{i+1}} H(b^{\mathbf{v}}, E^{\mathbf{v}}|\Pi) \\
&= \sum_{\mathbf{v}\in\mathcal{F}_{i+1}} (H(E^{\mathbf{v}}) + H(b^{\mathbf{v}}|\Pi, E^{\mathbf{v}})) \\
&\leq \sum_{\mathbf{v}\in\mathcal{F}_{i+1}} (1 + H(b^{\mathbf{v}}|\Pi, E^{\mathbf{v}})) \qquad \text{(since } E^{\mathbf{v}} \in \{0,1\}) \\
&= d + \sum_{\mathbf{v}\in\mathcal{F}_{i+1}} H(b^{\mathbf{v}}|\Pi, E^{\mathbf{v}} = 1)\cdot\mathbb{P}[E^{\mathbf{v}} = 1] + \sum_{\mathbf{v}\in\mathcal{F}_{i+1}} H(b^{\mathbf{v}}|\Pi, E^{\mathbf{v}} = 0)\cdot\mathbb{P}[E^{\mathbf{v}} = 0]
\end{aligned} \tag{38}$$

We now bound the terms on the rhs. First, since for every $\mathbf{v}$ one has that $b^{\mathbf{v}}$ has a fixed number of nonzeros in uniformly random positions by definition,

$$H(b^{\mathbf{v}}|\Pi, E^{\mathbf{v}} = 0) \leq H(b^{\mathbf{v}}). \tag{39}$$

We now bound the second sum on the last line of (38). First recall that if $E^{\mathbf{v}} = 1$, then $|M^{\mathbf{v}}_{ALG}| > cn/(6k)$, and $M^{\mathbf{v}}_{ALG}$ is a subset of the edges of $G'(u_{1:i-1}, \mathbf{v}; B_{1:i})$. Let $b := B_i$. Recall that for every vertex $y \in T_i \setminus T_i^{\mathbf{v}}$ we include edges incident on it in $H_i^{\mathbf{v}}$ if $b_y^{\mathbf{v}} = 1$ and do not otherwise. We thus have, for every $y \in T_i \setminus T_i^{\mathbf{v}}$ such that $\delta(y) \cap M^{\mathbf{v}}_{ALG} \neq \emptyset$ that $b_y^{\mathbf{v}} = 1$. Define for $\mathbf{v} \in \mathcal{F}_{i+1}$

$$\gamma^{\mathbf{v}} := \frac{|(T_i \setminus T_i^{\mathbf{v}}) \cap M^{\mathbf{v}}_{ALG}|}{|T_i \setminus T_i^{\mathbf{v}}|},$$

and note that whenever $|M^{\mathbf{v}}_{ALG}| > cn/(6k)$, we get using Lemma 19, **(1)** and **(4)**, as well as the fact that $\epsilon < 1/10$ by our choice of parameters (since $\epsilon \leq (c/k)^C$ for a sufficiently large constant $C \geq 1$),

$$\gamma^{\mathbf{v}} = \frac{cn/(6k)}{|T_i \setminus T_i^{\mathbf{v}}|} \geq c/30. \tag{40}$$

We may assume that $\gamma^{\mathbf{v}} \leq 1/3$ (remove some edges from $M^{\mathbf{v}}_{ALG}$ otherwise). We have

$$\begin{aligned}
H(b^{\mathbf{v}}|M^{\mathbf{v}}_{ALG}, E^{\mathbf{v}} = 1) &\leq \log_2 \binom{|T_i \setminus T_i^{\mathbf{v}}| - |M^{\mathbf{v}}_{ALG}|}{\lceil(1-\xi)|T_i \setminus T_i^{\mathbf{v}}|\rceil - |M^{\mathbf{v}}_{ALG}|} \\
&\leq \log_2 \binom{(1-\gamma^{\mathbf{v}})|T_i \setminus T_i^{\mathbf{v}}|}{(1-\xi-\gamma^{\mathbf{v}})|T_i \setminus T_i^{\mathbf{v}}|},
\end{aligned} \tag{41}$$

where we used the assumption that $\gamma^{\mathbf{v}} \leq 1/3$ and the fact that $\xi < 1/3$ by our setting of parameters. We thus get

$$H(b^{\mathbf{v}}|M^{\mathbf{v}}_{ALG}, E^{\mathbf{v}} = 1) \leq (1-\gamma^{\mathbf{v}})|T_i \setminus T_i^{\mathbf{v}}| H_2\left(1 - \frac{\xi}{1-\gamma^{\mathbf{v}}}\right),$$

since

$$\log_2 \binom{(1-\gamma^{\mathbf{v}})|T_i \setminus T_i^{\mathbf{v}}|}{(1-\xi-\gamma^{\mathbf{v}})|T_i \setminus T_i^{\mathbf{v}}|} = \log_2 \binom{(1-\gamma^{\mathbf{v}})|T_i \setminus T_i^{\mathbf{v}}|}{(1-\frac{\xi}{1-\gamma^{\mathbf{v}}})(1-\gamma^{\mathbf{v}})|T_i \setminus T_i^{\mathbf{v}}|}$$
$$\leq (1-\gamma^{\mathbf{v}})|T_i \setminus T_i^{\mathbf{v}}| H_2\left(1-\frac{\xi}{1-\gamma^{\mathbf{v}}}\right),$$

where the last transition is by subadditivity of entropy. At this point we also note that

$$\begin{aligned}(1-\gamma^{\mathbf{v}})H_2\left(1-\frac{\xi}{1-\gamma^{\mathbf{v}}}\right) &= \xi\log_2(1/\xi) + \xi\ln 2 - \xi\log\frac{1}{1-\gamma^{\mathbf{v}}} + O(\xi^2)\\ &\leq H_2(1-\xi) - \xi\log\frac{1}{1-\gamma^{\mathbf{v}}} + O(\xi^2)\\ &\leq H_2(1-\xi) - \Omega(\xi\cdot c) + O(\xi^2) \qquad \text{(by (40) and Claim 31).}\end{aligned}$$

since $H_2(1-\xi) = \xi\log_2(1/\xi) + \xi\ln 2 + O(\xi^2)$ and ξ is smaller than a constant. Putting the above bounds together, and noting that by subadditivity of entropy

$$H(b^{\mathbf{v}}) \leq |T_i \setminus T_i^{\mathbf{v}}| H_2(1-\xi) = |T_i \setminus T_i^{\mathbf{v}}| \cdot (\xi\log_2(1/\xi) + \xi\ln 2 + O(\xi^2)),$$

we get, since ξ is smaller than c by a large constant factor by our choice of ξ, that

$$\begin{aligned}H(b^{\mathbf{v}}|M_{ALG}, E^{\mathbf{v}}=1) &\leq H(b^{\mathbf{v}}) - \Omega(c\cdot\xi)\cdot|T_i \setminus T_i^{\mathbf{v}}|\\ &\leq H(b^{\mathbf{v}}) - \Omega(c\cdot\xi/k)\cdot|T|\end{aligned}$$

Using this upper bound in (38), we get

$$\begin{aligned}H(\mathbf{B}_i|\Pi) &\leq d + \sum_{\mathbf{v}\in\mathcal{F}_{i+1}} H(b^{\mathbf{v}}|\Pi, E^{\mathbf{v}}=1)\cdot\mathbb{P}[E^{\mathbf{v}}] + \sum_{\mathbf{v}\in\mathcal{F}_{i+1}} H(b^{\mathbf{v}})\cdot(1-\mathbb{P}[E^{\mathbf{v}}])\\ &\leq d + \sum_{\mathbf{v}\in\mathcal{F}_{i+1}} (H(b^{\mathbf{v}}) - \Omega(c\cdot\xi/k)\cdot|T|)\cdot\mathbb{P}[E^{\mathbf{v}}] + \sum_{\mathbf{v}\in\mathcal{F}_{i+1}} H(b^{\mathbf{v}})\cdot(1-\mathbb{P}[E^{\mathbf{v}}])\\ &\leq d + \sum_{\mathbf{v}\in\mathcal{F}_{i+1}} H(b^{\mathbf{v}}) - \Omega(c\cdot\xi/k)\cdot|T|\cdot\sum_{\mathbf{v}\in\mathcal{F}_{i+1}}\mathbb{P}[E^{\mathbf{v}}]\\ &= H(\mathbf{B}_i) - \Omega(c\cdot\xi/k)\cdot|T|\cdot|\mathcal{F}_{i+1}|\cdot\mathbb{E}_{\mathbf{v}\sim UNIF(\mathcal{F}_{i+1})}[E^{\mathbf{v}}] + d\\ &\leq H(\mathbf{B}_i) - \Omega(c\cdot\xi/k^2)\cdot|T|\cdot|\mathcal{F}_{i+1}| + d \qquad \text{(by (37))}\\ &\leq H(\mathbf{B}_i) - \Omega(c\cdot\xi/k^2)\cdot|T|\cdot|\mathcal{F}_{i+1}|\\ &\leq H(\mathbf{B}_i) - \Omega_k(c)\cdot dn.\end{aligned}$$

Using this bound in (36), we get $I(\mathbf{B}_i : \Pi) \geq \Omega_k(c)\cdot dn$, and therefore

$$s \geq H(\Pi) \geq I(\mathbf{B}_i;\Pi) \geq \Omega_{c,k}(nd),$$

as required. ■

Claim 31 *For every $\xi > 0$ the function $(1-\gamma)H_2(1-\frac{\xi}{1-\gamma})$ is decreasing in γ for all $\gamma \in (0, 1-\xi)$.*

Proof: We have

$$\begin{aligned}(1-\gamma)H_2\left(1-\frac{\xi}{1-\gamma}\right) &= \frac{1}{\ln 2}\cdot(1-\gamma)\left[\left(1-\frac{\xi}{1-\gamma}\right)\ln\frac{1}{1-\frac{\xi}{1-\gamma}}+\frac{\xi}{1-\gamma}\ln\frac{1-\gamma}{\xi}\right]\\ &= \frac{1}{\ln 2}\cdot\left[(1-\gamma-\xi)\ln\frac{1-\gamma}{1-\gamma-\xi}+\xi\ln\frac{1-\gamma}{\xi}\right]\\ &= \frac{1}{\ln 2}\cdot\left[(1-\gamma-\xi)\ln\left(1+\frac{\xi}{1-\gamma-\xi}\right)+\xi\ln\frac{1-\gamma}{\xi}\right]\end{aligned}$$

Since $\xi\ln\frac{1-\gamma}{\xi}$ is decreasing in $\gamma\in(0,1)$, it suffices to show that $(1-\gamma-\xi)\ln\left(1+\frac{\xi}{1-\gamma-\xi}\right)$ is decreasing in γ for $\gamma\in(0,1-\xi)$. Letting $x=1-\gamma-\xi$, it suffices to show that $x\ln(1+\frac{\xi}{x})$ is increasing in x for $x\in(0,1-\xi)$. Rescaling x by ξ, it suffices to show that $x\ln(1+\frac{1}{x})$ is increasing in x for all $x>0$. The derivative with respect to x is $\ln(1+\frac{1}{x})-\frac{1}{x+1}$, which approaches 0 as $x\to\infty$. The derivative of this function is $-\frac{1}{x(x+1)^2}$, which is negative for all $x>0$, and thus $\ln(1+\frac{1}{x})-\frac{1}{x+1}>0$ for all $x>0$. ■

We can now give

Proof of Theorem 1: Since ALG provides a better than $(1-1/e+c)$-approximation for some constant $c>0$ by assumption, there exists integer k such that $1-(1-1/k)^k+c/2\le 1-1/e+c$ (we assume that $2/c$ is an integer, which can be ensured by reducing c by at most a factor of 2).

Setting parameters. Let G' be generated as per Definition 26 with parameters selected as follows. First let $\xi=\eta=(c/k)^A$ for a sufficiently large integer $A>1$ (recall that ξ is the rate at which we subsample edges of G to obtain G'). Then let $\epsilon=(\eta/k)^{2B}$ for a sufficiently large integer $B>1$ (note that $1/\epsilon^{1/2}$ is an integer). Finally let m be an integer multiple of $1/\epsilon$, let $w=\epsilon m$ and let $W=w\cdot k/(\epsilon\cdot\theta)$, where $\theta=\eta$.

We have by Claim 27 that the graph G' (as per Definition 26) contains matching of size at least $(1-O(\xi+k^3\epsilon/\eta))|S|$ with probability at least $97/100$. We also note that

$$\begin{aligned}|S| &= \sum_{i=0}^{k-1}|S_i|+|S| = \sum_{i=0}^{k-1}(1-1/k)^i|Y|/k+(1-1/k)^k|Y|\pm O(k^4\epsilon/\theta)|Y|\\ &= (1\pm O(k^4\epsilon/\theta))|Y|\\ &= (1\pm c/100)|Y|\end{aligned}$$

by Lemma 19, **(1)** and **(2)**, since $O(k^4\epsilon/\theta)=O(k^4\epsilon/\eta)<c/100$ when A and B above are larger than an absolute constant (as we verify below in (43)). Thus, the algorithm must output a matching of size at least

$$(1-(1-1/k)^k+c/2)(1-c/100)^2|Y|\ge(1-(1-1/k)^k+c/4)|Y| \tag{42}$$

with probability at least $1/2$. The inequality above uses the fact that

$$\begin{aligned}O(\xi+k^4\epsilon/\eta) &= O((c/k)^A+k^4(\eta/k)^{2B}/\eta)\\ &= O((c/k)^A+(\eta/k)^{2B-4})\\ &= O((c/k)^A+(c/k)^{2B-4})\\ &< c/100\end{aligned} \tag{43}$$

as long as A and B are larger than an absolute constant.

Now let M_{ALG} be the matching output by a single pass streaming algorithm ALG on the graph G' presented in the order prescribed by our input distribution. For convenience we define M_{ALG} to be the empty set if ALG outputs an edge that was not in G'. By Lemma 28 we have

$$|M_{ALG}| \leq (1-1/k)^k |Y| + \sum_{i=0}^{k-1} |E(H_i^{\mathbf{u}_{i+1}}) \cap M_{ALG}| + O(k^3 \epsilon^{1/2}/\theta)|Y|. \tag{44}$$

Using this together with (44), as well as the fact that $O(k^3\epsilon^{1/2}/\theta) < c/100$ as long as A and B above are larger than an absolute constant, we get

$$\mathbb{P}\left[\sum_{i=0}^{k-1} |E(H_i^{\mathbf{u}_{i+1}}) \cap M_{ALG}| > (c/8)\cdot n\right] > 1/2$$

by assumption of the theorem. We now have by Lemma 30 that the space complexity s of the algorithm satisfies $s = \Omega_{c,k}(nd)$. Since $n = m^{4m}$ and $d = 2^{\Omega(\epsilon^2 m)} = 2^{\Omega(m)}$ for any fixed ϵ by Lemma 16, we get that $nd = n^{1+\Omega(1/\log\log n)}$, as required.

■

4 Multipass approximation for matchings

In this section we present our algorithm for approximating matchings in multiple passes in the vertex arrival setting, proving Theorem 3.

4.1 The algorithm

Let $G = (P, Q, E)$ denote a bipartite graph. We assume that vertices in P arrive in the stream together with all their edges. At each step the algorithm maintains a fractional matching $\{f_e\}_{e\in E}$, where the capacity of each vertex in Q is infinite and the capacity of each vertex $u \in P$ is equal to the number of times it has appeared in so far (i.e. always between 1 and k). The capacity of an edge $e = (u, v), u \in P, v \in Q$ is equal to the capacity of u. For a vertex $u \in P$ we write $\delta(u)$ to denote the set of neighbors of u in G.

The fractional matching f_e is initialized at zero, and upon arrival of a vertex $u \in P$ the algorithm continuously assigns a single unit of water to its least loaded neighbors. At the end of the k passes we obtain a bona-fide matching by reducing the load of vertices on the Q side that were assigned more than k units of fractional mass down to k units (simply reduce the load on neighboring edges). Scaling the resulting allocation by $1/k$ gives a feasible fractional matching, which can then be rounded to an integral matching using standard techniques in nearly linear time in the support size of the matching. The algorithm for processing a vertex $u \in P$ upon arrival is summarized in Algorithm 1 below.

Algorithm 1: PROCESSVERTEX(G, u, $\delta(u)$)

1: WATERFILLING($G', u, \delta(u)$) ▷ Assign one unit of water to least loaded neighbors
2: REMOVECYCLES(G', f).

The function WATERFILLING($G', u, \delta(u)$) increases the load of the least loaded neighbors of u simultaneously (with other neighbors joining if the load reaches their level) until one unit of water in total is dispensed out of u. Here the support of the fractional matching $\{f_e\}_{e\in E}$ maintained by the algorithm is denoted by G'. The function REMOVECYCLES(G', f) reroutes flow among cycles that could have emerged in the process, ensuring that the flow is supported on at most $|P| + |Q| - 1$ edges.

Efficient implementation. First note that WATERFILLING$(G', u, \delta(u))$ can be implemented to run in time $O(|\delta(u)| \log n)$. Indeed, we need to find θ such that

$$\sum_{e=(u,v)\in\delta(u)} \max\{\theta - c_v, 0\} = 1,$$

where c_v is the load of $v \in Q$ in the current fractional allocation. The function on the lhs is non-decreasing for all $\theta \geq \min_{e=(u,v)\in\delta(u)} c_v$, so the root can be found to within polynomial precision in $O(\log n)$ time using binary search.

Similarly, the function REMOVECYCLES can be implemented to run in nearly linear time at the expense of a loss of an $O(\log n)$ factor in space complexity. To achieve this we first buffer incoming vertices until the number of edges received is $\Theta(n)$ and only perform cycle removal after such a batch has been received. Let $f : E \to \mathbb{R}$ denote the allocation corresponding to one such batch. Write $f = \sum_{i=0}^{O(\log n)} 2^{-i} f_i$, where $f_i : E \to \{0, 1\}$ encode the sets of edges whose i-th bit in the allocation f is set to 1. Denote the corresponding edge sets by $E_i \subseteq E$, $i = 0, 1, \ldots, O(\log n)$. Now for every E_i run DFS to find cycles, and note that every time a cycle in E_i is found, we zero out half of the edges on the cycle while rerouting flow in f_i. Thus, the amount of work on E_i is indeed linear in its size, resulting in a nearly linear runtime bound overall.

We now turn to analyzing the approximation ratio. We first give a sketch of the proof under additional assumptions on the graph G, and then proceed to give the relevant definitions and the complete argument.

4.2 Analysis in a simple case (when G has a perfect matching)

In this section we assume that $G = (P, Q, E)$ has a perfect matching M in order to illustrate the main idea behind our analysis.

We start with

Definition 32 (Level sets b^k) *For each $k \geq 1$ and all $x \geq 0$ denote by $b^k(x)$ the number of vertices in Q that have load* at least *x after k passes in Algorithm 1.*

Note that $b^k(x)$ is non-increasing in x and $b^k(x) - b^{k-1}(x) \geq 0$ for all x. Furthermore, we have

$$b^k(0) = |M| \text{ and } \int_0^\infty b^k(x)dx = k|M|. \tag{45}$$

The first equality holds since G is assumed to contain a perfect matching, and the second holds since every vertex $u \in P$ contributed 1 unit of water, amounting to $|M| = |P|$ amount of water overall, and (45) calculates the sum of loads on all $v \in Q$. Furthermore, note that the size of the matching constructed by the algorithm after k passes is exactly equal to

$$\frac{1}{k} \int_0^k b^k(x)dx, \tag{46}$$

since every vertex $v \in Q$ with load x contributes $\frac{1}{k} \cdot \min\{k, x\}$ to the matching. Hence the approximation ratio after k passes is at least

$$1 - \frac{1}{|M|} \cdot \frac{1}{k} \int_k^\infty b^k(x)dx, \tag{47}$$

where we used (45) to convert (46) into (47). Thus, it is sufficient to lower bound $\int_0^k b^k(x)dx$ in order to analyze the approximation ratio, and we turn to bounding this quantity.

First consider the case $k = 1$. For each such vertex u consider its match $M(u)$. Since u ended up at level at least x after the first pass, its match $M(u)$ must be at level at least x after the first pass as well, as levels are

non-decreasing. Hence, we have

$$\begin{aligned} b^1(x) &= |\{u \in P : u \text{ is at level } \geq x \text{ after first pass}\}| \\ &\geq |\{u \in P : u \text{ allocated some water at level } \geq x \text{ during first pass}\}| \\ &\geq \int_x^\infty b^1(s)ds \end{aligned} \tag{48}$$

for all $x \geq 0$. This, however, together with (45) can be shown to imply that $\int_x^\infty b^1(s)ds \leq |M| \cdot e^{-x}$ for all x. We thus get using (47) that the approximation ratio after one pass is at least $1 - 1/e$.

Now suppose that $k > 1$ and consider vertices $v \in Q$ that are at level at least x after k-th pass, but were at a lower level after $(k-1)$-th pass. There are exactly $b^k(x) - b^{k-1}(x)$ such vertices. Since these vertices u were at level at least x after k-th pass, their matches $M(u)$ must have also been at level at least x after the k-th pass, implying similarly to the above that

$$b^k(x) \geq \int_x^\infty (b^k(s) - b^{k-1}(s))ds \tag{49}$$

for all $x \geq 0$. The above equation implies that for all $k \geq 1$

$$\int_x^\infty b^k(s)ds \leq |M| \cdot \int_x^\infty F^k(s)ds, \tag{50}$$

where $1-F^k(x)$ is the cdf of the Gamma distribution with scale 1 and shape k, i.e. $F^k(x) = \int_x^\infty e^{-s}s^{k-1}/(k-1)!ds$. Using this in (47) yields the desired bound on the approximation ratio, i.e. $1 - e^{-k}k^{k-1}/k!$.

4.3 Analysis in a general case

The proof sketch we gave in the previous subsection works under the assumption that G has a perfect matching. The general case is more involved. While the analysis above proceeds by showing that not too much mass will be in the tail $\int_k^\infty b^k(x)dx$, here we find it more convenient to show that substantial mass will be in the head of the distribution, i.e. bound $\int_0^k b^k(x)dx$ from below. We extend the argument using a careful reweighting of vertices and scaling of levels guided by the structure of the *canonical decomposition* of G introduced in [GKK12], which we now define.

Let $G = (P, Q, E)$ denote a bipartite graph. For a set $S \subseteq P$ we denote the set of neighbors of S by $\Gamma(S)$. For a number $\alpha > 0$ the graph G is said to have vertex expansion at least α if $|\Gamma(S)| \geq \alpha|S|$ for all $S \subseteq P$. The canonical decomposition of G is defined as follows:

Definition 33 (Canonical decomposition) *Let $G = (P, Q, E)$ denote a bipartite graph. A partition of $Q = \bigcup_{j\in\mathcal{I}} T_j, T_j \cap T_i = \emptyset, j \neq i$ and $P = \bigcup_{j\in\mathcal{I}} S_j, S_j \cap S_i = \emptyset, j \neq i$ together with numbers $\alpha_j > 0$, where $\alpha_j \leq 1$ for $j \leq 0$ and $\alpha_j > 1$ for $j > 0$ is called a* canonical partition *if*

1. *for all i one has $\Gamma\left(\bigcup_{j\in\mathcal{I}, j\leq i} S_j\right) \subseteq \bigcup_{j\in\mathcal{I}, j\leq i} T_j$;*
2. *$|\Gamma(S) \cap T_j| \geq \alpha_j|S|$ for all $S \subseteq S_j$ for all $j \in \mathcal{I}$;*
3. *$|T_j|/|S_j| = \alpha_j$, for all $j \in \mathcal{I}$.*

Here $\mathcal{I} \subset \mathbb{Z}$ is a set of indices.

See Fig. 1 for an illustration.

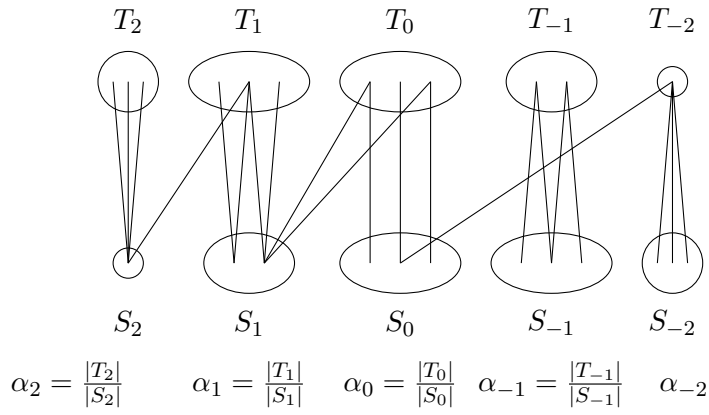

Figure 1: Canonical decomposition of a bipartite graph. Note that edges from S_i only go to T_j with $j \leq i$ (property (1)).

Vertex capacities and canonical matching. First, define vertex capacities as follows. For $u \in P$ let j be such that $u \in S_j$(see Fig. 1), and let $c(u) := \min\{1, \alpha_j\}$. Similarly, for $v \in Q$ let j be such that $v \in T_j$ (see Fig. 1) and let $c(v) := \min\{1, 1/\alpha_j\}$. We will also use

Claim 34 (Monotonicity of capacities) *For every $i \leq j$ and every $v \in T_i, w \in T_j$ one has $c(v) \geq c(w)$. Similarly, for every $i \leq j$ and every $v \in S_i, w \in S_j$ one has $c(v) \leq c(w)$.*

Proof: Follows by monotonicity of α_j's. ■

Definition 35 (Canonical matching) *Let $M : E \to [0, 1]$ be a (possibly fractional) matching in G such that $\sum_{e \in \delta(u)} x_e = c(u)$ for all $u \in P = \bigcup_j S_j$ and $\sum_{e \in \delta(v)} x_e = c(v)$ for all $v \in Q = \bigcup_j T_j$.*

Such a matching exists by properties (2) and (3) of the canonical decomposition. Furthermore, any such M is a maximum matching in G, since $|M| = |C|$, where $C = \left(\cup_{j:\alpha_j \geq 1} S_j\right) \bigcup \left(\cup_{j:\alpha_j < 1} T_j\right)$ forms a vertex cover in G by property (1) of the canonical decomposition. For every integer $j = 1, \ldots, k$ and $e \in E$ we let $\widetilde{M}_e^j \in [0, 1]$ denote the load assigned by our algorithm in the j-th pass to edge e. Note that $\widetilde{M}$ does not necessarily form a matching, but for every $u \in P$ and every j one has $\sum_{e=\delta(u)} \widetilde{M}^j(e) = 1$, since every vertex on the P side dispenses one unit of water in every pass. We note that

Claim 36 *For every graph G, if (S_j, T_j) is the canonical decomposition (as per Definition 33), M a canonical matching in G (as per Definition 35), and vertex capacities as defined above, then $\sum_{u \in P} c(u) = \sum_{v \in Q} c(v) = |M|$.*

Shadow allocation and density function $\phi_v^k(x)$. We will use the concept of a *shadow allocation*, in which whenever a units of water are added to a vertex $v \in Q$ in the original allocation, $a/c(v)$ units of water are added to v in the shadow allocation. Now whenever water from a vertex $u \in P$ is added to vertex $v \in Q$ at level x during the j-th pass *in the shadow allocation*, we let $\phi_v^j(x) := c(u)$, where ϕ is the *density function*.

The following claim is crucial for our analysis:

Claim 37 *For every graph G, if M is a maximum matching in G, vertex capacities c and density function ϕ are defined as above, one has $\sum_{v \in Q} c(v) \int_0^\infty \phi_v^j(x) dx = |M|$ for all $j = 1, \ldots, k$.*

Proof:

$$\begin{aligned}\sum_{v\in Q} c(v)\int_0^\infty \phi_v^j(x)dx &= \sum_{v\in Q} c(v) \sum_{e=(u,v)\in\delta(v)} c(u)\cdot \widetilde{M}^j(e)/c(v)\\ &= \sum_{v\in Q}\sum_{e=(u,v)\in\delta(v)} c(u)\cdot \widetilde{M}^j(e)\\ &= \sum_{u\in P} c(u) \sum_{e=(u,v)\in\delta(v)} \widetilde{M}^j(e)\\ &= \sum_{u\in P} c(u)\\ &= |M|,\end{aligned}$$

where the first equality is by definition of the shadow allocation, the fourth is by definition of $\widetilde{M}^j$ and the last is by Claim 36. ■

Load of a vertex and level of an edge. The core of our analysis will consist of bounding the distribution of water levels among vertices in Q in the shadow allocation, showing that there cannot be too many highly overloaded vertices. For a vertex $v \in Q$ let $l^k(v)$ denote the load of v in the shadow allocation after the k-th pass. For an edge $e = (u, v)$ let $l^k(e)$ denote the load of v in the shadow allocation after u is processed in the k-th pass. The key property of $l^k(e)$ that we need is given by

Lemma 38 *For every $k \geq 1$, every $e = (u, v)$ such that $\widetilde{M}^k(e) > 0$ and $f = (u, w)$ such that $M(f) > 0$ one has $l^k(f) \geq l^k(e)$.*

Proof: Denote the load of v in the original (as opposed to shadow) allocation after u is processed during the k-th pass by x, and denote the load of w in the original (as opposed to shadow) allocation after u is processed during the k-th pass by y. We have $y \geq x$ by the definition of the waterfilling algorithm. Also note that $l^k(e) = x/c(v)$ and $l^k(f) = y/c(w)$ by definition of the shadow allocation. By the properties of the canonical decomposition one has $v \in T_i, w \in T_j$ for some $i \leq j$, and hence $c(v) \geq c(w)$ by Claim 34. We therefore have

$$l^k(f) = y/c(w) \geq y/c(v) \geq x/c(v) = l^k(e),$$

as required. ■

Reweighted level set sizes b^k. For every $x \geq 0$, integer $k \geq 1$ we let $b^k(x)$ denote the (weighted) number of vertices with load at least x in the shadow allocation, defined as follows:

$$b^k(x) = \sum_{v\in Q} c(v)\cdot \mathbb{1}_{l^k(v)\geq x}.$$

Note that $b^k(0) = \sum_{v\in Q} c(v) = |M|$ for every k. We have

Lemma 39 *Algorithm 1 constructs a matching of size at least $\frac{1}{k}\int_0^k b^k(x)dx$.*

Proof: For a vertex $v \in Q$ let $l^k_{org}(v)$ denote the water level of at v in the original allocation after k passes. Then v contributes $\frac{1}{k}\min\{k, l_{org}(v)\}$ to the matching. At the same time $l^k(v) = l^k_{org}(v)/c(v)$, so

$$\begin{aligned}
\frac{1}{k}\int_0^k b^k(x)dx &= \frac{1}{k}\int_0^k \sum_{v\in Q} c(v)\cdot \mathbb{1}_{l^k(v)\geq x} dx \\
&= \frac{1}{k}\sum_{v\in Q} c(v)\cdot \min\{k, l^k(v)\} \\
&= \frac{1}{k}\sum_{v\in Q} c(v)\cdot \min\{k, l^k_{org}(v)/c(v)\} \\
&= \frac{1}{k}\sum_{v\in Q} \min\{c(v)\cdot k, l^k_{org}(v)\} \\
&\leq \sum_{v\in Q}\frac{1}{k}\min\{k, l^k_{org}(v)\},
\end{aligned}$$

where we used the fact that $c(v) \leq 1$ for all v in the last step. This completes the proof of the lemma. ■

Bounding the evolution of $b^k(x)$. In what follows we derive bounds on the reweighted level set sizes $b^k(x)$, which then allow us to lower bound $\frac{1}{k}\int_0^k b^k(x)dx$. We start with

Lemma 40 *One has for all $x \geq 0$ and all $k \geq 1$*

$$b^k(x) \geq \int_x^\infty \sum_{v\in Q} c(v)\phi^k_v(s)ds.$$

Proof: First note that

$$\begin{aligned}
\int_x^\infty \sum_{v\in Q} c(v)\cdot\phi^k_v(s)ds &= \sum_{v\in Q} c(v)\cdot \sum_{\substack{e=(u,v)\in\delta(v)\\ l^k(e)\geq x}} c(u)\cdot \widetilde{M}^k(e)/c(v) \\
&= \sum_{v\in Q}\sum_{\substack{e=(u,v)\in\delta(v)\\ l^k(e)\geq x}} c(u)\cdot\widetilde{M}^k(e) \\
&= \sum_{u\in P} c(u) \sum_{\substack{e=(u,v)\in\delta(u)\\ l^k(e)\geq x}} \widetilde{M}^k(e)
\end{aligned} \tag{51}$$

by definition of the shadow allocation and density function ϕ. Recall that for an edge $e = (u, v)$ we let $l^k(e)$ denote the load of vertex v right after u arrives in the k-th pass. At the same time,

$$\begin{aligned}
b^k(x) &= \sum_{v\in Q} c(v)\cdot \mathbb{1}_{l^k(v)\geq x} \\
&= \sum_{\substack{v\in Q\\ l^k(v)\geq x}} \sum_{e=(u,v)\in E} M(e) \qquad \text{(since } \sum_{e=(u,v)\in E} M(e) = c(v) \text{ for every } v\text{)} \\
&\geq \sum_{u\in P}\sum_{\substack{e=(u,v)\in E\\ l^k(e)\geq x}} M(e).
\end{aligned} \tag{52}$$

Now recall that by Lemma 38 for every $u \in P$ if u dispensed some water at level at least x *in the shadow allocation* during the k-th pass, i.e. if

$$\sum_{e=(u,v)\in\delta(u):l^k(e)\geq x} \widetilde{M}^k(e) > 0,$$

then its canonical matches, namely vertices v such that $M_{(u,v)} > 0$, were at level at least x in the shadow allocation after u was processed during k-th pass. In particular, in that case we have

$$\sum_{e=(u,v)\in E:l^k(e)\geq x} M(e) = c(u).$$

Since $\sum_{e=(u,v)\in\delta(u):l^k(e)\geq x} \widetilde{M}^k(e) \leq 1$ always, we thus get for every $u \in P$

$$c(u) \sum_{e=(u,v)\in\delta(u):l^k(e)\geq x} \widetilde{M}^k(e) \leq \sum_{e=(u,v)\in E:l^k(v)\geq x} M(e).$$

Indeed, if the sum on the lhs is positive, then the sum on the rhs equals $c(u)$ (which suffices since the lhs is bounded by $c(u)$), and if the sum in the lhs is zero, then the inequality holds trivially since the rhs is nonnegative. Summing over $u \in P$, we get

$$\sum_{u\in P} c(u) \sum_{e=(u,v)\in\delta(u):l^k(e)\geq x} \widetilde{M}^k(e) \leq \sum_{u\in P} \sum_{e=(u,v)\in E:l^k(e)\geq x} M(e).$$

This, together with (51) and (52) yields $b^k(x) \geq \int_x^\infty \sum_{v\in Q} c(v) \cdot \phi_v^k(s)ds$, as required.

■

We now get, letting $b^0 \equiv 0$ for convenience,

Lemma 41 *For all $x \geq 0$ and all $k \geq 1$ one has $|M| - b^k(x) \leq \int_0^x (b^k(s) - b^{k-1}(s))ds$.*

Proof: By Lemma 40 we have $b^k(x) \geq \int_x^\infty \sum_{v\in Q} c(v)\phi_v^k(s)ds$. Putting this together with Claim 37 we get $|M| - b^k(x) \leq \int_0^x \sum_{v\in Q} c(v)\phi_v^k(s)ds$ for all $x \geq 0$ and $k \geq 1$. To complete the proof, we note that, since $\phi_v^k(s) \leq 1$ for all v, k, s,

$$\begin{aligned}
\int_0^x \sum_{v\in Q} c(v)\phi_v^k(s)ds &\leq \int_0^x \sum_{v\in Q} c(v) \cdot \mathbb{1}[v \text{ is allocated water at level } s \text{ in pass } k]ds \\
&= \int_0^x \sum_{v\in Q} c(v) \cdot (\mathbb{1}_{l^k(v)\geq s} - \mathbb{1}_{l^{k-1}(v)<s})ds \\
&= \int_0^x (b^k(s) - b^{k-1}(s))ds
\end{aligned}$$

for all $k \geq 1$ and $x \geq 0$, where we let $b^0 \equiv 0$ for convenience. ■

We now prove lower bounds on $b^k(x)$. Recall that for integer $k \geq 1$

$$F^k(x) = \int_x^\infty e^{-s}s^{k-1}/(k-1)!ds = \sum_{i=0}^{k-1} e^{-x}x^i/i!, \tag{53}$$

so that $1 - F^k(x)$ is the cdf of the Gamma distribution with scale 1 and shape k. We now prove our main lower bound on b^k:

Lemma 42 *For every $k \geq 1$ for all $x \geq 0$ one has $\int_0^x b^k(s)ds \geq |M| \cdot \int_0^x F^k(s)ds$.*

Proof: We prove the claim of the lemma by induction on k.

Base: $k = 1$ Recall that by Lemma 41 one has

$$|M| - b^1(x) \leq \int_0^x b^1(s)ds. \tag{54}$$

Letting $f(x) = \int_0^x b^1(s)ds$, we get by rearranging (54) and noting that $f'(x) = b^1(x)$ that $f'(x) \geq |M| - f(x)$ for all $x \geq 0$. We also have $f(0) = 0$. Thus, we have $f(x) \geq |M| \cdot (1 - e^{-x}) = |M| \cdot \int_0^x F^1(s)ds$, as required.

Inductive step: $k - 1 \to k$ We need to prove that

$$\int_0^x b^k(s)ds \geq |M| \cdot \int_0^x F^k(s)ds. \tag{55}$$

Using Lemma 41 and the inductive hypothesis, we get for all $x \geq 0$

$$\begin{aligned} b^k(x) &\geq |M| - \int_0^x (b^k(s) - b^{k-1}(s))ds \\ &= |M| - \int_0^x b^k(s)ds + \int_0^x b^{k-1}(s)ds \\ &\geq |M| - \int_0^x b^k(s)ds + |M| \cdot \int_0^x F^{k-1}(s)ds. \qquad \text{(by the inductive hypothesis)} \end{aligned} \tag{56}$$

Let $f(x) = \int_0^x b^k(s)ds$ (note that $f(0) = 0$). We have from (56) that

$$f'(x) \geq |M| - f(x) + |M| \cdot \int_0^x F^{k-1}(s)ds.$$

Thus, for all $x \geq 0$ one has $f(x) \geq g(x)$, where $g(x)$ is given by the solution of

$$g'(x) = |M| - g(x) + |M| \cdot \int_0^x F^{k-1}(s)ds,$$

which we now solve. The latter implication holds by Claim 50 applied to $|M| - f(x)$. Note that since $g(0) = 0$, we have by the above that $g'(0) = |M|$. Thus, $h(x) = g'(x)$ satisfies

$$h'(x) = -h(x) + |M| \cdot F^{k-1}(x), h(0) = |M|. \tag{57}$$

The solution to (57) is given by

$$h(x) = |M| \cdot e^{-x} \left(\int_0^x e^s F^{k-1}(s)ds + 1 \right). \tag{58}$$

Calculating the integral in (58) using the expression for $F^{k-1}(s)$ given by (53) yields

$$\int_0^x e^s F^{k-1}(s)ds = \int_0^x e^s \int_s^\infty \frac{1}{(k-2)!} z^{k-2} e^{-z} dz ds = \int_0^x \sum_{j=0}^{k-2} \frac{1}{j!} s^j ds = \sum_{j=1}^{k-1} \frac{1}{j!} x^j, \tag{59}$$

and hence

$$h(x) = |M| \cdot e^{-x} \left(\int_0^x e^s F^{k-1}(s) ds + 1 \right) = |M| \cdot e^{-x} \left(\sum_{j=1}^{k-1} \frac{1}{j!} x^j + 1 \right) = |M| \cdot F^k(x)$$

by (53). We thus get $g(x) = \int_0^x h(s)ds = |M| \cdot \int_0^x F^k(s)ds$, and therefore $\int_0^x b^k(s)ds \geq f(x) \geq |M| \cdot \int_0^x F^k(s)ds$ as required.

■

Given Lemma 42, we immediately obtain

Theorem 43 *Algorithm 1 achieves a* $(1 - e^{-k} \frac{k^{k-1}}{(k-1)!})$*-approximation to maximum matchings in* k *passes over the input stream.*

Proof: By Lemma 39 together with Lemma 42 the approximation ratio is at least

$$\frac{1}{|M|} \cdot \frac{1}{k} \int_0^k b^k(x)dx \geq \frac{1}{k} \int_0^k F^k(x)dx = 1 - \frac{1}{k} \int_k^\infty F^k(x)dx.$$

We now recall (by (53)) that $F^k(x) = \sum_{j=0}^{k-1} e^{-x} x^j / j!$. Integrating by parts, we have

$$\int_k^\infty e^{-x} x^j / j! dx = -e^{-x} x^j / j! \Big|_k^\infty + \int_k^\infty e^{-x} x^{j-1} / (j-1)! dx,$$

and hence

$$\frac{1}{k} \int_k^\infty F^k(x)dx = \frac{1}{k} \int_k^\infty \sum_{j=0}^{k-1} e^{-x} x^j / j! dx = \frac{1}{k} \sum_{j=0}^{k-1} (k-j) e^{-k} k^j / j!.$$

Since

$$\frac{1}{k} \sum_{j=0}^{k-1} (k-j) e^{-k} k^j / j! = \sum_{j=0}^{k-1} e^{-k} k^j / j! - \sum_{j=1}^{k-1} e^{-k} k^{j-1} / (j-1)! = e^{-k} k^{k-1} / (k-1)!,$$

we thus get

$$\frac{1}{|M|} \cdot \frac{1}{k} \int_0^k b^k(x)dx \geq 1 - e^{-k} k^{k-1} / (k-1)!,$$

as required. ■

Remark 44 *We note that the approximation ratio satisfies* $\frac{e^{-k} k^{k-1}}{(k-1)!} = \frac{1}{\sqrt{2\pi k}} + O(k^{-3/2})$.

5 Gap-existence

In this section we show how our techniques yield an efficient algorithm for Gap-existence, thereby proving Theorem 4.

We now describe DISCRETIZEDWATERFILLING, which is a version of Algorithm 1. We will explicitly maintain a subset $I^* \subset I$ of size $O(|A|/\epsilon)$ while relying on an oracle NEWNEIGHBOR(a, I^*) that, given any set $I^* \subseteq I$, outputs any node $i \in I \setminus I^*$ that a is connected to or $\emptyset$ if all neighbors of a are in I^*. The difference

Algorithm 2: DISCRETIZEDWATERFILLING(G, a, ϵ, k)

$I^* \leftarrow \emptyset$ ▷$N(a) \subseteq I$ stands for the vertex neighborhood of $a \in A$
while ≤ 1 unit of water allocated **do**
  while $\exists\, i \in N(a) \cap I^*$ with level $< (\epsilon/4)k$ and ≤ 1 unit of water has been allocated **do**
    Allocate water to i until it is at level $(\epsilon/4)k$
    if one unit of water has been allocated from a **then**
      return
    end if
    $i \leftarrow$ NEWNEIGHBOR(a, I^*) ▷NEWNEIGHBOR(a, I^*) returns $\emptyset$ if all neighbors of a are in I^*
    if $i \neq \emptyset$ **then**
      $I^* \leftarrow I^* \cup \{i\}$
    else
      break from both loops
    end if
  end while
end while
Perform water filling on neighbors in I^*.

First we prove

Lemma 45 *The space used by Algorithm 2 is $O(|A|/\epsilon)$.*

Proof: Call a vertex *saturated* if the amount of water in it is at least ϵk. The number of saturated vertices is $O(|A|/\epsilon)$ since there are $k|A|$ units of water in the system, and each saturated vertex accounts for at least ϵk. We say that an unsaturated vertex i belongs to $a \in A$ if i was added to I^* when NEWNEIGHBOR was called from a. Note that for each $a \in A$ only one $i \in I$ belongs to a. Thus, this amounts to at most $|A|$ additional vertices. ■

Our algorithm for Gap-Existence is as follows:

Algorithm 3: GAPEXISTENCE(G, ϵ)

Run DISCRETIZEDWATERFILLING(G) with $k = O(\log(\sum_{a \in A} B_a/\epsilon)/\epsilon^2)$.
Output **YES** if at most $\epsilon/2$ water is allocated above level $k/(1-\epsilon/2)$, and **NO** otherwise.

We now assume that we are in the **YES** case, i.e. there exists a matching with budgets B_a, and prove that the algorithm will find a matching with budgets $\lfloor (1-\epsilon)B_a \rfloor$.

We recall definitions of levels and level set sizes below.

Definition 46 *Define $l^k(i)$ to be the level of water at vertex i after the k-th pass (here we refer to the level of water in the actual allocation constructed by waterfilling, not the shadow allocation used for analysis purposes in Section 4.3).*

Definition 47 (Level set sizes) *For each $k \geq 1$ and all $x \geq 0$ denote by $b^k(x)$ the number of vertices in I that have load* at least *x after k passes, i.e. the number of vertices $i \in I$ with $l^k(i) \geq x$.*

Note that $b^k(x)$ is non-increasing in x and $b^k(x) - b^{k-1}(x) \geq 0$ for all x, and $\int_0^\infty b^k(x)dx = k|A|$, since every vertex dispenses one unit of water in every pass.

We now note that the allocation constructed by DISCRETIZEDWATERFILLING can be used to obtain a matching as follows: we first scale the allocation by a factor of $1-\epsilon/2$, then take all water allocated below level k. Dividing by $1/k$ gives a matching where every vertex in A is assigned at least

$$(1-\epsilon/2)\cdot B_a - \frac{1}{k}\int_{k/(1-\epsilon/2)}^{\infty} b^k(x)dx \tag{60}$$

fractional mass. Thus, if the second term is bounded by $\epsilon/2$, then the graph contains a matching with budgets $\lfloor(1-\epsilon)B_a\rfloor, a\in A$, i.e. if the algorithm outputs **YES**, it is correct. In what follows we show that in the **YES** case, i.e. when the input graph admits a matching with budgets $B_a, a\in A$, the second term is indeed bounded by $\epsilon/2$.

For simplicity of notation we assume from now on that every $a\in A$ is replaced with B_a unit demand copies (and we use A to denote the set of those copies, abusing notation somewhat). We assume that we are in the **YES** case, i.e. the original graph contains a matching with budgets B_a, and thus the new graph admits a perfect matching of the A side – denote this matching by M. For every edge e of G, every k we let $\widetilde{M}^k(e)$ denote the amount of fractional mass allocated along edge e in the k-th pass. For an edge $e=(a,i)$ we let $l^k(e)$ denote the load of vertex i right after a arrives in the k-th pass, and let $l^k(i)$ denote the load of i after the k-th pass.

Lemma 48 *One has for all $k\geq 1$ and $x\geq(\epsilon/4)\cdot k$*

$$b^k(x)\geq\int_x^{\infty}(b^k(s)-b^{k-1}(s))ds. \tag{61}$$

where $b^0\equiv 0$.

Proof: Intuitively, the lemma follows since if a vertex $a\in A$ ended up allocating water at level at least x during the k-th pass, its match must have been at level at least x when a arrived. Together with the fact that levels are monotone increasing this gives the result. We now give the details.

First note that

$$\begin{aligned}\int_x^{\infty}\sum_{i\in I}(b^k(s)-b^{k-1}(s))ds &= \sum_{i\in I}\sum_{\substack{e=(a,i)\in\delta(i)\\ l^k(e)\geq x}}\widetilde{M}^k(e)\\ &= \sum_{a\in A}\sum_{\substack{e=(a,i)\in\delta(a)\\ l^k(e)\geq x}}\widetilde{M}^k(e)\end{aligned} \tag{62}$$

At the same time

$$\begin{aligned}b^k(x) &= \sum_{i\in I}\mathbb{1}_{l^k(i)\geq x}\\ &\geq \sum_{\substack{i\in I\\ l^k(i)\geq x}}\sum_{e=(a,i)\in E}M(e)\qquad\left(\text{since }\sum_{e=(a,i)\in E}M(e)\leq 1\text{ for every }i\in I\right)\\ &\geq \sum_{a\in A}\sum_{\substack{e=(a,i)\in E\\ l^k(e)\geq x}}M(e).\end{aligned} \tag{63}$$

Now note that if $a\in A$ dispensed some water at level at least x during the k-th pass, i.e. if

$$\sum_{e=(a,i)\in\delta(a):l^k(e)\geq x}\widetilde{M}^k(e)>0,$$

then vertices i such that $M_{(a,i)} > 0$ were at level at least x after a was processed during k-th pass. In particular, in that case we have

$$\sum_{e=(a,i)\in E: l^k(e)\geq x} M(e) = 1.$$

Since $\sum_{e=(a,i)\in\delta(a): l^k(e)\geq x} \widetilde{M}^k(e) \leq 1$ always, we thus get for every $a \in A$

$$\sum_{e=(a,i)\in\delta(a): l^k(e)\geq x} \widetilde{M}^k(e) \leq \sum_{e=(a,i)\in E: l^k(i)\geq x} M(e).$$

Indeed, if the sum on the lhs is positive, then the sum on the rhs equals 1 (which suffices since the lhs is bounded by 1), and if the sum in the lhs is zero, then the inequality holds trivially since the rhs is nonnegative. Summing over $a \in A$, we get

$$\sum_{a\in A}\sum_{e=(a,i)\in\delta(a): l^k(e)\geq x} \widetilde{M}^k(e) \leq \sum_{a\in A}\sum_{e=(a,i)\in E: l^k(e)\geq x} M(e).$$

This, together with (62) and (63) yields $b^k(x) \geq \int_x^\infty \sum_{i\in I}(b^k(s) - b^{k-1}(s))ds$, as required. ■

We now get, letting $\Delta = (\epsilon/4)\cdot k$ to simplify notation,

Lemma 49 *For all $k \geq 1$ and all $x \geq \Delta$, then*

$$\int_x^\infty b^k(s)ds \leq |A| \cdot \int_{x-\Delta}^\infty F^k(s)ds. \tag{64}$$

Proof: We prove the lemma by induction on k.

Base: $k = 1$ Recall that by Lemma 48 one has

$$b^1(x) \geq \int_x^\infty b^1(s)ds, \tag{65}$$

for all $x \geq \Delta$. Let $f(x) = \int_x^\infty b^1(s)ds$, so that $f(x) \leq |A|$ for every x, and note that for $x \geq \Delta$

$$f'(x) = -b^1(x) \leq -\int_x^\infty b^1(s)ds = -f(x).$$

Let $g(x)$ be a function such that $g(\Delta) = |A|$ and $g'(x) = -g(x)$ for all $x \geq \Delta$. Then we have $f(x) \leq g(x)$ for all $x \geq \Delta$, and therefore for all $x \geq \Delta$

$$\int_x^\infty b^1(s)ds = f(x) \leq g(x) = |A|e^{-x+\Delta} = |A| \cdot \int_{x-\Delta}^\infty e^{-s}ds = |A| \cdot \int_{x-\Delta}^\infty F^1(s)ds.$$

Inductive step: $k-1 \to k$ We need to prove that

$$\int_x^\infty b^k(s)ds \leq |A| \cdot \int_x^\infty F^k(s)ds. \tag{66}$$

Using Lemma 48 we get for all $x \geq \Delta$

$$b^k(x) \geq \int_x^\infty (b^k(s) - b^{k-1}(s))ds = \int_x^\infty b^k(s)ds - |A| \cdot \int_{x-\Delta}^\infty F^{k-1}(x), \tag{67}$$

where we used the inductive hypothesis to upper bound $\int_x^\infty b^{k-1}(s)ds$ with $|A| \cdot \int_{x-\Delta}^\infty F^{k-1}(s)ds$. We thus have that the function $f(x) = \int_x^\infty b^k(s)ds$ satisfies

$$f'(x) \leq -f(x) + |A| \cdot \int_{x-\Delta}^\infty F^{k-1}(x), f(\Delta) \leq k|A|,$$

where the last condition comes from the fact that every vertex in A dispenses one unit of water in every pass overall, so the total amount of water dispensed at level x or above is bounded by $k|A|$.

Let g satisfy

$$g'(x) = -g(x) + |A| \cdot \int_{x-\Delta}^\infty F^{k-1}(s)ds, g(\Delta) = k|A|, \tag{68}$$

so that $g(x) \geq f(x) = \int_x^\infty b^k(s)ds$ for $x \geq \Delta$ (by Claim 50 below). Let $h(x) = g'(x)$, so that

$$h'(x) = -h(x) - |A| \cdot F^{k-1}(x - \Delta) \tag{69}$$

and $h(\Delta) = g'(\Delta) = -g(\Delta) + |A| \cdot \int_0^\infty F^{k-1}(s)ds = -k|A| + (k-1)|A| = -|A|$. The second to last equality holds since $\int_0^\infty F^{k-1}(s)ds$ equals the expectation of the sum of $k-1$ exponentially distributed variables of unit scale, which is $k-1$.

The solution to (69) is given by

$$h(x) = e^{-x+\Delta}\left(-|A| \cdot \int_0^{x-\Delta} e^s F^{k-1}(s)ds - |A|\right). \tag{70}$$

Calculating the integral in (70) yields

$$\begin{aligned}
\int_0^{x-\Delta} e^s F^{k-1}(s)ds &= \int_0^{x-\Delta} e^s \int_s^\infty \frac{1}{(k-2)!} z^{k-2} e^{-z} dz ds \\
&= \int_0^{x-\Delta} \sum_{j=0}^{k-2} \frac{1}{j!} s^j ds \\
&= \sum_{j=1}^{k-1} \frac{1}{j!}(x-\Delta)^j,
\end{aligned} \tag{71}$$

and hence

$$\begin{aligned}
h(x) &= e^{-x+\Delta}\left(-|A| \cdot \sum_{j=1}^{k-1} \frac{1}{j!}(x-\Delta)^j - |A|\right) \\
&= e^{-x+\Delta}\left(-|A| \cdot \sum_{j=0}^{k-1} \frac{1}{j!}(x-\Delta)^j\right) \\
&= -|A| \cdot F^k(x-\Delta)
\end{aligned}$$

by (53). Therefore

$$\begin{aligned}
g(x) &= g(\Delta) + \int_\Delta^x h(s)ds \\
&= k|A| + \int_\Delta^x h(s)ds \\
&= -\int_x^\infty h(s)ds \\
&= |A| \cdot \int_{x-\Delta}^\infty F^k(s-\Delta)ds,
\end{aligned}$$

and $\int_x^\infty b^k(s)ds = f(x) \le g(x) = |A| \cdot \int_{x-\Delta}^\infty F^k(s-\Delta)ds$, as required.

■

Claim 50 *For every $g : \mathbb{R} \to \mathbb{R}$, if $f : \mathbb{R} \to \mathbb{R}$ satisfies $f'(x) \le -f(x)+g(x)$, $f(0) = a$ for some $a = 0$, then $f(x) \le h(x)$ for $h : \mathbb{R} \to \mathbb{R}$ that satisfies $h(x) = -h'(x) + g(x), h(0) = a$ and is pointwise non-decreasing in a.*

Proof: Let $q(x) = e^x f(x)$, so that

$$q'(x) = e^x f(x) + e^x f'(x) \le e^x f(x) + e^x(-f(x) + g(x)) = e^x g(x).$$

Integrating from 0 to x, we get $q(x) \le q(0) + \int_0^x e^s g(s)ds = f(0) + \int_0^x e^s g(s)ds$. Letting $h(x) := e^{-x}(\int_0^x e^s g(s)ds + f(0))$, we note that

$$f(x) = e^{-x}q(x) \le e^{-x}(f(0) + \int_0^x e^s g(s)ds) = h(x).$$

It remains to note that $h'(x) = -h(x) + g(x)$ for all $x \ge 0$, $h(0) = f(0) = a$, and $h(x)$ is non-decreasing in $f(0) = a$, as required. ■

We will need

Lemma 51 *For all $k \ge 1$ and $\delta \ge 0$*

$$\frac{1}{k}\int_{k(1+\delta)}^\infty F^k(x)dx \le k \cdot e^{-\delta k}(1+\delta)^k$$

Proof: Recalling that $F^k(x) = \sum_{j=0}^{k-1} e^{-x}x^j/j!$ and using integration by parts

$$\int_{k(1+\delta)}^\infty e^{-x}x^j/j!dx = -e^{-x}x^j/j!\big|_{k(1+\delta)}^\infty + \int_{k(1+\delta)}^\infty e^{-x}x^{j-1}/(j-1)!dx,$$

we get

$$\begin{aligned}
\int_{k(1+\delta)}^\infty F^k(x)dx &= \int_{k(1+\delta)}^\infty \sum_{j=0}^{k-1} e^{-x}x^j/j!dx \\
&= \sum_{j=0}^{k-1}(k-j)e^{-k(1+\delta)}(k(1+\delta))^j/j! \\
&\le e^{-\delta k}(1+\delta)^k \cdot e^{-k}\sum_{j=0}^{k-1} k^{j+1}/j! \\
&\le e^{-\delta k}(1+\delta)^k \cdot ke^{-k}\sum_{j=0}^{\infty} k^j/j! \\
&= k \cdot e^{-\delta k}(1+\delta)^k.
\end{aligned} \tag{72}$$

■

We now use Lemma 49 to upper bound the second term in (60) by $\epsilon/2$, as required. By Lemma 49 we have

$$\frac{1}{k}\int_{k/(1-\epsilon/2)}^\infty b^k(x)dx \le |A| \cdot F^k(k/(1-\epsilon/2) - (\epsilon/4)k). \tag{73}$$

Since

$$\begin{aligned} k/(1-\epsilon/2) - (\epsilon/4)k &= k \cdot \frac{1-\epsilon/2+\epsilon/2}{1-\epsilon/2} - k \cdot \frac{(\epsilon/4)(1-\epsilon/2)}{1-\epsilon/2} \\ &= k\left(1 + \frac{\epsilon/2 - (\epsilon/4)(1-\epsilon/2)}{1-\epsilon/2}\right) \\ &\geq k\,(1+\epsilon/4) \end{aligned}$$

when ϵ is smaller than an absolute constant, we get, letting $\delta = \epsilon/4$ for convenience of notation, that by Lemma 51

$$\frac{1}{k}\int_{k(1+\delta)}^{\infty} b^k(x)dx \leq k \cdot e^{-k(\delta - \ln(1+\delta))} \leq k \cdot e^{-\Omega(\epsilon^2 k)} \tag{74}$$

as long as $\delta = \Theta(\epsilon)$ is smaller than an absolute constant. Hence, letting $k = C\ln(\frac{1}{\epsilon}\cdot\sum_{a\in A} B_a)/\epsilon^2$ for a sufficiently large constant $C > 0$, we get by (60) that every advertizer is satisfied with budget at least

$$\begin{aligned} (1-\epsilon/2)\cdot B_a - \frac{1}{k}\int_{k/(1-\epsilon/2)}^{\infty} b^k(x)dx &\geq \alpha \cdot B_a - \left(\sum_{a\in A} B_a\right)\cdot k e^{-\Omega(\epsilon^2 k)} \\ &\geq (1-\epsilon/2)\cdot B_a - \epsilon/2 \\ &\geq (1-\epsilon)\cdot B_a. \end{aligned}$$

This completes the proof of Theorem 4.

References

[AB21] Sepehr Assadi and Soheil Behnezhad. Beating two-thirds for random-order streaming matching. *CoRR*, abs/2102.07011, 2021.

[ABB+19] Sepehr Assadi, MohammadHossein Bateni, Aaron Bernstein, Vahab S. Mirrokni, and Cliff Stein. Coresets meet EDCS: algorithms for matching and vertex cover on massive graphs. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 1616–1635. SIAM, 2019.

[AG11] Kook Jin Ahn and Sudipto Guha. Linear programming in the semi-streaming model with application to the maximum matching problem. In Luca Aceto, Monika Henzinger, and Jirí Sgall, editors, *Automata, Languages and Programming - 38th International Colloquium, ICALP 2011, Zurich, Switzerland, July 4-8, 2011, Proceedings, Part II*, volume 6756 of *Lecture Notes in Computer Science*, pages 526–538. Springer, 2011.

[ALT21] Sepehr Assadi, Cliff Liu, and Robert Tarjan. An auction algorithm for bipartite matching in streaming and massively parallel computation models. *SOSA*, 2021.

[Ber20] Aaron Bernstein. Improved bounds for matching in random-order streams. In Artur Czumaj, Anuj Dawar, and Emanuela Merelli, editors, *47th International Colloquium on Automata, Languages, and Programming, ICALP 2020, July 8-11, 2020, Saarbrücken, Germany (Virtual Conference)*, volume 168 of *LIPIcs*, pages 12:1–12:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

[CCD+10] Denis Xavier Charles, Max Chickering, Nikhil R. Devanur, Kamal Jain, and Manan Sanghi. Fast algorithms for finding matchings in lopsided bipartite graphs with applications to display ads. In David C. Parkes, Chrysanthos Dellarocas, and Moshe Tennenholtz, editors, *Proceedings 11th ACM Conference on Electronic Commerce (EC-2010), Cambridge, Massachusetts, USA, June 7-11, 2010*, pages 121–128. ACM, 2010.

[EKS09] Sebastian Eggert, Lasse Kliemann, and Anand Srivastav. Bipartite graph matchings in the semi-streaming model. In Amos Fiat and Peter Sanders, editors, *Algorithms - ESA 2009, 17th Annual European Symposium, Copenhagen, Denmark, September 7-9, 2009. Proceedings*, volume 5757 of *Lecture Notes in Computer Science*, pages 492–503. Springer, 2009.

[ELSW13] Leah Epstein, Asaf Levin, Danny Segev, and Oren Weimann. Improved bounds for online preemptive matching. In Natacha Portier and Thomas Wilke, editors, *30th International Symposium on Theoretical Aspects of Computer Science, STACS 2013, February 27 - March 2, 2013, Kiel, Germany*, volume 20 of *LIPIcs*, pages 389–399. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2013.

[FKM+04] Joan Feigenbaum, Sampath Kannan, Andrew McGregor, Siddharth Suri, and Jian Zhang. On graph problems in a semi-streaming model. In Josep Díaz, Juhani Karhumäki, Arto Lepistö, and Donald Sannella, editors, *Automata, Languages and Programming: 31st International Colloquium, ICALP 2004, Turku, Finland, July 12-16, 2004. Proceedings*, volume 3142 of *Lecture Notes in Computer Science*, pages 531–543. Springer, 2004.

[FKM+05] Joan Feigenbaum, Sampath Kannan, Andrew McGregor, Siddharth Suri, and Jian Zhang. Graph distances in the streaming model: the value of space. In *Proceedings of the Sixteenth An-*

nual ACM-SIAM Symposium on Discrete Algorithms, SODA 2005, Vancouver, British Columbia, Canada, January 23-25, 2005, pages 745–754. SIAM, 2005.

[FLN+02] Eldar Fischer, Eric Lehman, Ilan Newman, Sofya Raskhodnikova, Ronitt Rubinfeld, and Alex Samorodnitsky. Monotonicity testing over general poset domains. In John H. Reif, editor, *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 474–483. ACM, 2002.

[GKK12] Ashish Goel, Michael Kapralov, and Sanjeev Khanna. On the communication and streaming complexity of maximum bipartite matching. In Yuval Rabani, editor, *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012*, pages 468–485. SIAM, 2012.

[GKM+19] Buddhima Gamlath, Michael Kapralov, Andreas Maggiori, Ola Svensson, and David Wajc. Online matching with general arrivals. In David Zuckerman, editor, *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pages 26–37. IEEE Computer Society, 2019.

[Kap21] Michael Kapralov. Space lower bounds for approximating maximum matching in the edge arrival model. In Dániel Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pages 1874–1893. SIAM, 2021.

[KMM12] Christian Konrad, Frédéric Magniez, and Claire Mathieu. Maximum matching in semi-streaming with few passes. In Anupam Gupta, Klaus Jansen, José D. P. Rolim, and Rocco A. Servedio, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 15th International Workshop, APPROX 2012, and 16th International Workshop, RANDOM 2012, Cambridge, MA, USA, August 15-17, 2012. Proceedings*, volume 7408 of *Lecture Notes in Computer Science*, pages 231–242. Springer, 2012.

[KMT11] Chinmay Karande, Aranyak Mehta, and Pushkar Tripathi. Online bipartite matching with unknown distributions. In Lance Fortnow and Salil P. Vadhan, editors, *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 587–596. ACM, 2011.

[KVV90] Richard M. Karp, Umesh V. Vazirani, and Vijay V. Vazirani. An optimal algorithm for on-line bipartite matching. In Harriet Ortiz, editor, *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing, May 13-17, 1990, Baltimore, Maryland, USA*, pages 352–358. ACM, 1990.

[McG05] Andrew McGregor. Finding graph matchings in data streams. In Chandra Chekuri, Klaus Jansen, José D. P. Rolim, and Luca Trevisan, editors, *Approximation, Randomization and Combinatorial Optimization, Algorithms and Techniques, 8th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems, APPROX 2005 and 9th InternationalWorkshop on Randomization and Computation, RANDOM 2005, Berkeley, CA, USA, August 22-24, 2005, Proceedings*, volume 3624 of *Lecture Notes in Computer Science*, pages 170–181. Springer, 2005.

[MY11] Mohammad Mahdian and Qiqi Yan. Online bipartite matching with random arrivals: an approach based on strongly factor-revealing lps. In Lance Fortnow and Salil P. Vadhan, editors, *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 597–606. ACM, 2011.

[WW15] Yajun Wang and Sam Chiu-wai Wong. Two-sided online bipartite matching and vertex cover: Beating the greedy algorithm. In Magnús M. Halldórsson, Kazuo Iwama, Naoki Kobayashi, and Bettina Speckmann, editors, *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part I*, volume 9134 of *Lecture Notes in Computer Science*, pages 1070–1081. Springer, 2015.

A Proofs omitted from Section 3

Lemma 18 *(Restated) For every $m \geq 2$, integer $W \geq 1$ and $\delta' \in (0,1)$ such that $1/\delta'$ is an integer, if $Y = [m^4]^m$ and the set $\mathcal{S}$ is defined by*

$$\mathcal{S} = \{y \in Y : (y, \mathbf{u}) + \Delta_{\mathbf{u}} \mod W \in [a_{\mathbf{u}}, b_{\mathbf{u}}) \cdot W, \text{ for all } \mathbf{u} \in \mathcal{U}\},$$

where $\mathcal{U}$ is a collection of binary vectors of fixed length w and $a_{\mathbf{u}}, b_{\mathbf{u}} \in [0,1]$ are constant integer multiples of $1/L$ for an integer L, the following conditions hold if W is an integer multiple of $w \cdot lcm(L, 1/\delta')$, $\Delta_{\mathbf{u}}/W$ are multiples of $1/L$ and m is sufficiently large.

If $\max_{\substack{\mathbf{u}\in\mathcal{U},\mathbf{v}\in\mathcal{U}\\ \mathbf{u}\neq\mathbf{v}}}(\mathbf{u},\mathbf{v})/|\mathbf{v}| \leq \delta'$, then

$$\left| |\mathcal{S}| - |Y| \cdot \prod_{\mathbf{u}\in\mathcal{U}} (b_{\mathbf{u}} - a_{\mathbf{u}}) \right| \leq |\mathcal{U}|^2 (6L\delta' + 4/m) \cdot |Y|.$$

Before proving the lemma we introduce some definitions. Throughout this section we use the notation $Y = [m^4]^m$ for integer m. First define

Definition 52 (Bad vertices) *We let $B \subseteq Y$ denote the set of* bad *vertices, i.e. vertices with at least one coordinate close to 0 or m^4:*

$$B := \{x \in Y : \exists i \in [m] \text{ such that } x_i < m^2 \text{ or } x_i > m^4 - m^2\}.$$

We will use

Lemma 53 (The hypercube $Y = [m^4]^m$ contains few bad vertices) *For every integer $m > 1$, if $Y = [m^4]^m$, then $|B| \leq (2/m)|Y|$.*

Proof: Follows directly by a union bound

$$|B| \leq \sum_{i\in[m]} |\{x \in Y : x_i < m^2 \text{ or } x_i > m^4 - m^2\}| \leq m \cdot (2m^2/m^4) \cdot |Y| \leq (2/m)|Y|.$$

■

We will extensively use the notion of a discretization of the cube $Y = [m^4]^m$:

Definition 54 (Discretization with precision L) *For every integer $m, W, L > 1$, $\delta \in (0,1)$, every collection $\mathcal{U}$ of binary vectors of length m, every $q \in [L]^{\mathcal{U}}$ define*

$$S(q) := \left\{ y \in Y : (y, \mathbf{u}) \mod W \in \left[\frac{q_{\mathbf{u}} - 1}{L}, \frac{q_{\mathbf{u}}}{L} \right) \cdot W, \text{ for all } \mathbf{u} \in \mathcal{U} \right\},$$

and

$$\mathit{Int}_\delta(S(q)) := \left\{ y \in Y : (y, \mathbf{u}) \mod W \in \left[\frac{q_{\mathbf{u}} - 1}{L} + \delta, \frac{q_{\mathbf{u}}}{L} - \delta \right] \cdot W, \text{ for all } \mathbf{u} \in \mathcal{U} \right\}.$$

We will also use

Definition 55 (Shifting map ψ) *For every integer $m, W, L > 1$, every collection $\mathcal{U}$ of binary vectors of weight w such that W/w is an integer, for every pair $q, r \in [L]^{\mathcal{U}}$ let*

$$\psi_{r\to q}(y) := y + \sum_{\mathbf{u}\in\mathcal{U}} \frac{W}{L\cdot w}(q-r)_{\mathbf{u}} \cdot \mathbf{u}.$$

We will use

Lemma 56 *For every $\delta \in (0,1)$, integer $m, W, L > 1$, every collection $\mathcal{U}$ of binary vectors of weight w, if $Y = [m^4]^m$, $B \subseteq Y$ is the set of bad vertices (as per Definition 52), then the following conditions hold. If $|\mathcal{U}| \cdot (W/w) < m^2$, $\max_{\substack{\mathbf{u}\in\mathcal{U},\mathbf{v}\in\mathcal{U}\\ \mathbf{u}\neq\mathbf{v}}}(\mathbf{u},\mathbf{v})/|\mathbf{v}| \le \delta'$ for some $\delta' \in (0, \delta/|\mathcal{U}|)$, then for every pair $q, r \in [L]^{\mathcal{U}}$ we have*

$$\psi_{r\to q}(\mathit{Int}_\delta(S(r)) \setminus B) \subseteq S(q).$$

Proof: First note that for every $y \in \mathrm{Int}_\delta(S(r)) \setminus B$ one has for every $q \in [L]^{\mathcal{U}}$

$$\psi_{r\to q}(y) = y + \sum_{\mathbf{u}\in\mathcal{U}} (q-r)_{\mathbf{u}} \cdot \mathbf{u} \cdot \frac{W}{L\cdot w} \in Y,$$

since

$$\begin{aligned}\left\|\sum_{\mathbf{u}\in\mathcal{U}} (q-r)_{\mathbf{u}} \cdot \mathbf{u} \cdot \frac{W}{L\cdot w}\right\|_\infty &\le \sum_{\mathbf{u}\in\mathcal{U}} \left\|(q-r)_{\mathbf{u}} \cdot \mathbf{u} \cdot \frac{W}{L\cdot w}\right\|_\infty \\ &\le |\mathcal{U}|\frac{W}{L\cdot w}||q-r||_\infty||\mathbf{u}||_\infty \le |\mathcal{U}|\cdot(W/w) < m^2.\end{aligned}$$

To obtain the last inequality we used the assumption that $\mathbf{u}$ is a binary vector, and the assumption of the lemma that $|\mathcal{U}|(W/w) < m^2$.

The rest of the proof proceeds in two steps. We first prove basic bounds on the dot product of $\psi_{r\to q}(y)$ with vectors $\mathbf{u} \in \mathcal{U}$, and then put these bounds together to obtain the result of the lemma. We have for every $y \in Y$ and $\mathbf{u} \in \mathcal{U}$

$$\begin{aligned}(\psi_{r\to q}(y), \mathbf{u}) &= (y,\mathbf{u}) + (q-r)_{\mathbf{u}} \cdot |\mathbf{u}| \cdot \frac{W}{L\cdot w} && \text{(since } \mathbf{u} \text{ is a binary vector)} \\ &\quad + \sum_{\mathbf{v}\in\mathcal{U},\mathbf{v}\neq\mathbf{u}} (q-r)_{\mathbf{v}} \cdot (\mathbf{v},\mathbf{u}) \cdot \frac{W}{L\cdot w} \\ &= (y,\mathbf{u}) + \frac{(q-r)_{\mathbf{u}}}{L} \cdot W && \text{(intended shift in direction of } \mathbf{u}) \\ &\quad + \sum_{\mathbf{v}\in\mathcal{U},\mathbf{v}\neq\mathbf{u}} (q-r)_{\mathbf{v}} \cdot (\mathbf{v},\mathbf{u}) \cdot \frac{W}{L\cdot w} && \text{(small error term from near-orthogonality)}\end{aligned} \tag{75}$$

We now bound the error term (the last line) in the previous equation. We have, using the assumption that $\max_{\substack{\mathbf{u}\in\mathcal{U},\mathbf{v}\in\mathcal{U}\\ \mathbf{u}\neq\mathbf{v}}}(\mathbf{u},\mathbf{v})/|\mathbf{v}| \le \delta'$ as well as the assumption that all vectors in $\mathcal{U}$ have the same Hamming weight

w, that

$$\begin{aligned}\left|\sum_{\mathbf{v}\in\mathcal{U},\mathbf{v}\neq\mathbf{u}}(q-r)_{\mathbf{v}}(\mathbf{v},\mathbf{u})\cdot\frac{W}{L\cdot w}\right| &\leq \sum_{\mathbf{v}\in\mathcal{U},\mathbf{v}\neq\mathbf{u}}|(q-r)_{\mathbf{v}}|\cdot\delta'\cdot|\mathbf{u}|\cdot\frac{W}{L\cdot w}\\ &\leq \sum_{\mathbf{v}\in\mathcal{U},\mathbf{v}\neq\mathbf{u}}|(q-r)_{\mathbf{v}}|\,\delta'\frac{W}{L}\\ &\leq \sum_{\mathbf{v}\in\mathcal{U},\mathbf{v}\neq\mathbf{u}}\delta' W \quad \text{(since } ||q-r||_\infty\leq L)\\ &\leq |\mathcal{U}|\delta' W\end{aligned} \tag{76}$$

Combining (75) and (76), we thus get

$$\left|(\psi_{r\to q}(y),\mathbf{u})-((y,\mathbf{u})+\frac{(q-r)_{\mathbf{u}}}{L}\cdot W)\right|\leq|\mathcal{U}|\delta' W. \tag{77}$$

Equipped with the bound above, we now proceed to complete the proof of the lemma.

We now show that for every $y\in \text{Int}_\delta(S(r))\setminus B$ one has $\psi_{r\to q}(y)\in S(q)$. Indeed, for each $y\in\text{Int}_\delta(S(r))$ one has by definition of $\text{Int}_\delta(S(q))$ (Definition 54)

$$\begin{aligned}((y,\mathbf{u})+\frac{(q-r)_{\mathbf{u}}}{L}\cdot W)\mod W &\in\left(\left[\frac{r_{\mathbf{u}}-1}{L}+\delta,\frac{r_{\mathbf{u}}}{L}-\delta\right]+\frac{(q-r)_{\mathbf{u}}}{L}\cdot W\right)\mod W\\ &\in\left[\frac{q_{\mathbf{u}}-1}{L}+\delta,\frac{q_{\mathbf{u}}}{L}-\delta\right]\mod W\end{aligned}$$

Combining the equation above with (77), we thus get for every $y\in Y\setminus B$

$$(\psi_{r\to q}(y),\mathbf{u})\mod W\in\left[\frac{q_{\mathbf{u}}-1}{L},\frac{q_{\mathbf{u}}}{L}\right)\mod W$$

since $\delta'<\delta/|\mathcal{U}|$ by assumption of the lemma. We have thus proved that for every $q,r\in[L]^{\mathcal{U}}$ one has

$$\psi_{r\to q}(\text{Int}_\delta(S(r))\setminus B)\subseteq S(q),$$

as required.

■

We will also use

Lemma 57 *For integer $m,w,W,L>1$ such that $m^2>W/w$, every vector $\mathbf{u}\in\{0,1\}^m$ of Hamming weight w, if $Y=[m^4]^m$, $B\subseteq Y$ is the set of bad vertices (as per Definition 52), then the following conditions hold. If $W/(L\cdot w)$ is a positive integer, then for every $\mathcal{I}\subseteq[L]$ we have*

$$\left|\left\{y\in Y:(y,\mathbf{u})\mod W\in\left[\frac{j-1}{L},\frac{j}{L}\right]\cdot W,j\in\mathcal{I}\right\}\right|\leq(|\mathcal{I}|/L+2/m)|Y|.$$

Proof: Consider a discretization of the cube (similarly to Definition 54) with $\mathcal{U}=\{\mathbf{u}\}$. Specifically, for $j\in[L]$ let

$$Z_j:=\left\{y\in Y:(y,\mathbf{u})\mod W\in\left[\frac{j-1}{L},\frac{j}{L}\right]\cdot W\right\}$$

and

$$Z'_j:=\left\{y\in Y\setminus B:(y,\mathbf{u})\mod W\in\left[\frac{j-1}{L},\frac{j}{L}\right]\cdot W\right\}.$$

We also let $z_j := |Z_j|$ and $z'_j := |Z'_j|$ for every $j \in [L]$.

To bound the size of Z_j and Z'_j, we use the shifting map ψ from Definition 55 with $\mathcal{U} = \{\mathbf{u}\}$, so that for every $i, j \in [L]$

$$\psi_{j\to i}(y) = y + (i-j)\cdot \mathbf{u} \cdot \frac{W}{L\cdot|\mathbf{u}|}.$$

Note that for every $y \in Y$ one has[1]

$$(\psi_{j\to i}(y), \mathbf{u}) = (y, \mathbf{u}) + (i-j)\cdot|\mathbf{u}|\cdot\frac{W}{L\cdot|\mathbf{u}|} = (y,\mathbf{u}) + \frac{i-j}{L}\cdot W,$$

and for every coordinate $s \in [m]$

$$(\psi_{j\to i}(y))_s = \left(y + (i-j)\cdot\mathbf{u}\cdot\frac{W}{L\cdot|\mathbf{u}|}\right)_s = y_s + (i-j)\cdot\mathbf{u}_s\cdot\frac{W}{L\cdot|\mathbf{u}|}.$$

Since for every $y \in Y \setminus B$ and every coordinate $s \in [m]$ we have $y_s \in [m^2, m^4 - m^2]$, $|i-j| \le L$ and $\mathbf{u}$ is a binary vector, we get that

$$y_s + (i-j)\cdot\mathbf{u}_s\cdot\frac{W}{L\cdot|\mathbf{u}|} \in [m^2 - W/|\mathbf{u}|, m^4 - m^2 + W/|\mathbf{u}|] \subseteq [m^4]$$

since $m^2 > W/|\mathbf{u}|$ by assumption of the lemma, as required. We thus conclude that for every $i, j \in [L]$ one has

$$\psi_{j\to i}(Z(j)\setminus B) \subseteq Z(i).$$

Since $\psi_{j\to i}$ is injective for all $i, j \in [L]$, we therefore have that $z'_j = |Z'_j| \le |Z_i| = z_i$ for all $i, j \in [L]$. We thus have, for any subset $\mathcal{I} \subseteq [L]$ of indices

$$\begin{aligned}
\sum_{j\in\mathcal{I}} z_j &\le \sum_{j\in\mathcal{I}} z'_j + \sum_{j\in\mathcal{I}}(z_j - z'_j) \\
&\le (|\mathcal{I}|/L)\sum_{j\in[L]} z_j + \sum_{j\in\mathcal{I}}(z_j - z'_j) && \text{(since } z'_j \le z_i \text{ for all } i\in[L]\text{)} \\
&\le (|\mathcal{I}|/L)\sum_{j\in[L]} z_j + |B| && \text{(since } Z_j \setminus Z'_j \subseteq B \text{ and } Z_j\text{'s are disjoint)} \\
&\le (|\mathcal{I}|/L)\sum_{j\in[L]} z_j + (2/m)|Y| && \text{(by Claim 52)}
\end{aligned}$$

as required. ■

Lemma 58 *For every $\delta \in (0,1)$ such that $1/\delta$ is an integer, integer $m, W, L > 1$ such that $W/(lcm(L, 1/\delta)\cdot w)$ is an integer, every vector $\mathbf{u} \in \{0,1\}^m$ of weight w, if $Y = [m^4]^m$, $B \subseteq Y$ is the set of bad vertices (as per Definition 52), then for every $\mathcal{I} \subseteq [L]$ we have*

$$\sum_{q\in[L]^{\mathcal{U}}} |S(q) \setminus Int_\delta(S(q))| \le |\mathcal{U}|(3\delta L + 2/m)\cdot|Y|$$

[1]Note that here we prove stronger properties of the shifting map than those proved in Lemma 56, but only for the special case of $\mathcal{U}$ containing a single element.

Proof: One has using Definition 54

$$\begin{aligned}&\sum_{q\in[L]^{\mathcal{U}}}|S(q)\setminus \mathrm{Int}_\delta(S(q))|\\ \leq&|\mathcal{U}|\max_{\mathbf{u}\in\mathcal{U}}\left\{y\in Y:(y,\mathbf{u})\mod W\in\left[\frac{q}{L}-\delta,\frac{q}{L}+\delta\right]\cdot W,\quad\text{for some } q\in[L]\right\},\end{aligned}\tag{78}$$

Let L' be the least integer multiple of $1/\delta$ and L. By Lemma 57 with parameter L' (note that the preconditions as satisfied since $W/(\mathrm{lcm}(L,1/\delta)\cdot w)$ is an integer by assumption of the lemma) and

$$\begin{aligned}\mathcal{I}&:=\left\{q'\in[L']:\frac{q'}{L'}\in\left[\frac{q}{L}-\delta,\frac{q}{L}+\delta\right]\text{ for some } q\in[L]\right\}\\&=\left\{q'\in[L']:q'\in\left[q\cdot\frac{L'}{L}-\delta\cdot L',q\cdot\frac{L'}{L}+\delta L'\right]\text{ for some } q\in[L]\right\},\end{aligned}$$

we get, using the fact that $|\mathcal{I}|\leq(2\delta L'+1)\cdot L$, that

$$\begin{aligned}&\left|\left\{y\in Y:(y,\mathbf{u})\mod W\in\left[\frac{j-1}{L'},\frac{j}{L'}\right]\cdot W,j\in\mathcal{I}\right\}\right|\\&\leq(|\mathcal{I}|/L'+2/m)|Y|\\&\leq((2\delta+1/L')L+2/m)|Y|\\&\leq(3\delta L+2/m)|Y|\quad(\text{since } L'\geq 1/\delta)\end{aligned}$$

Putting this together with (78) yields the result.

■

Proof of Lemma 18: Consider a discretization of the cube with parameters L and $\delta\in(0,1)$ (see Definition 54). We use $\delta=2|\mathcal{U}|\cdot\delta'$. Let $A_{\mathbf{u}}=a_{\mathbf{u}}\cdot L, B_{\mathbf{u}}=b_{\mathbf{u}}\cdot L,\mathbf{u}\in\mathcal{U}$ be integers such that $a_{\mathbf{u}}=A_{\mathbf{u}}/L, b_{\mathbf{u}}=B_{\mathbf{u}}/L$. Recall that per Definition 54 we have

$$S(q)=\left\{y\in Y:(y,\mathbf{u})\mod W\in\left[\frac{q_{\mathbf{u}}-1}{L},\frac{q_{\mathbf{u}}}{L}\right)\cdot W,\text{ for all }\mathbf{u}\in\mathcal{U}\right\},$$

and

$$\mathrm{Int}_\delta(S(q))=\left\{y\in Y:(y,\mathbf{u})\mod W\in\left[\frac{q_{\mathbf{u}}-1}{L}+\delta,\frac{q_{\mathbf{u}}}{L}-\delta\right]\cdot W,\text{ for all }\mathbf{u}\in\mathcal{U}\right\}.$$

We let

$$\begin{aligned}\mathcal{J}&:=\left\{q\in[L]^{\mathcal{U}}:((q_{\mathbf{u}}/L)\cdot W+\Delta_{\mathbf{u}})\mod W\in[A_{\mathbf{u}}/L,B_{\mathbf{u}}/L)\cdot W\right\}\\&=\left\{q\in[L]^{\mathcal{U}}:((q_{\mathbf{u}}/L+r_{\mathbf{u}}/L)\cdot W)\mod W\in[A_{\mathbf{u}}/L,B_{\mathbf{u}}/L)\cdot W\right\},\end{aligned}$$

where $r_{\mathbf{u}}:=\Delta_{\mathbf{u}}\cdot L,\mathbf{u}\in\mathcal{U}$, are integers by assumption of the lemma. Note that

$$\mathcal{S}=\bigcup_{q\in\mathcal{J}}S(q),$$

and hence, since $S(q)\cap S(q')=\emptyset$ for $q\neq q'$, we have

$$|\mathcal{S}|=\sum_{q\in\mathcal{J}}|S(q)|.$$

Also note that $|\mathcal{J}|=\prod_{\mathbf{u}\in\mathcal{U}}(B_{\mathbf{u}}-A_{\mathbf{u}})=L^{|\mathcal{U}|}\cdot\prod_{\mathbf{u}\in\mathcal{U}}(b_{\mathbf{u}}-a_{\mathbf{u}})$.

The proof proceeds in two steps. We first lower bound the size of $\mathcal{S}$ and then upper bound it. The arguments are quite similar, and rely on technical lemmas derived in the rest of this section.

Lower bound. First, by Lemma 56 that for every $q, r \in [L]^{\mathcal{U}}$ one has

$$\psi_{q\to r}(\text{Int}_\delta(S(q)) \setminus B) \subseteq S(r). \tag{79}$$

Note that the preconditions of the lemma are satisfied since $|\mathcal{U}| \cdot (W/w) < m^2$ (m is sufficiently large as function of other parameters) and we set $\delta = 2|\mathcal{U}| \cdot \delta'$. Applying (79) for every $q \in [L]^{\mathcal{U}}$ and $r \in \mathcal{J}$ and noting that the mapping $\psi_{q\to r}$ is injective gives

$$|\mathcal{J}| \cdot \sum_{q\in[L]^{\mathcal{U}}} |\text{Int}_\delta(S(q)) \setminus B| \le L^{|\mathcal{U}|} \sum_{q\in\mathcal{J}} |S(q)|.$$

We thus get

$$\begin{aligned}
\sum_{q\in\mathcal{J}} |S(q)| &\ge \left(\prod_{\mathbf{u}\in\mathcal{U}} (b_{\mathbf{u}} - a_{\mathbf{u}})\right) \cdot \sum_{q\in[L]^{\mathcal{U}}} |\text{Int}_\delta(S(q)) \setminus B| \\
&\ge \left(\prod_{\mathbf{u}\in\mathcal{U}} (b_{\mathbf{u}} - a_{\mathbf{u}})\right) \cdot \sum_{q\in[L]^{\mathcal{U}}} (|S(q) \setminus B| - |S(q) \setminus \text{Int}_\delta(S(q))|) \\
&\ge \left(\prod_{\mathbf{u}\in\mathcal{U}} (b_{\mathbf{u}} - a_{\mathbf{u}})\right) \cdot (|Y| - |B| - \sum_{q\in[L]^{\mathcal{U}}} |S(q) \setminus \text{Int}_\delta(S(q))|) \\
&\ge \left(\prod_{\mathbf{u}\in\mathcal{U}} (b_{\mathbf{u}} - a_{\mathbf{u}})\right) \cdot |Y| - |\mathcal{U}|(3\delta L + 4/m)|Y|. \quad \text{(by Lemma 58 and Lemma 53)}
\end{aligned} \tag{80}$$

We used Lemma 58 and the fact that $\prod_{\mathbf{u}\in\mathcal{U}}(b_{\mathbf{u}} - a_{\mathbf{u}}) \le 1$ since $a_{\mathbf{u}}, b_{\mathbf{u}} \in [0, 1]$ by assumption of the lemma to go from line 3 to line 4, and Lemma 53 to go from line 4 to line 5.

Upper bound. At the same time we also get, using again that by Lemma 56 that for every $q, r \in [L]^{\mathcal{U}}$ one has

$$\psi_{q\to r}(\text{Int}_\delta(S(q)) \setminus B) \subseteq S(r),$$

that

$$|\mathcal{J}| \cdot \sum_{q\in[L]^{\mathcal{U}}} |S(q)| \ge L^{|\mathcal{U}|} \sum_{q\in\mathcal{J}} |\text{Int}_\delta(S(q)) \setminus B|.$$

The above bound follows by noting that for every $q \in \mathcal{J}$ and $r \in [L]^{\mathcal{U}}$ one has $\psi_{q\to r}(\text{Int}_\delta(S(q)) \setminus B) \subseteq S(r)$, and the mapping $\psi_{q\to r}$ is injective. We thus get

$$\sum_{q\in\mathcal{J}} |\text{Int}_\delta(S(q)) \setminus B| \le \left(\prod_{\mathbf{u}\in\mathcal{U}} (b_{\mathbf{u}} - a_{\mathbf{u}})\right) \cdot \sum_{q\in[L]^{\mathcal{U}}} |S(q)| = \left(\prod_{\mathbf{u}\in\mathcal{U}} (b_{\mathbf{u}} - a_{\mathbf{u}})\right) \cdot |Y|. \tag{81}$$

We also have by Lemma 58 (applied with δ') and Lemma 53

$$\begin{aligned}
\sum_{q\in\mathcal{J}} |\text{Int}_\delta(S(q)) \setminus B| &\ge \sum_{q\in\mathcal{J}} |S(q)| - \sum_{q\in[L]^{\mathcal{U}}} |S(q) \setminus \text{Int}_\delta(S(q))| - |B| \\
&\ge \sum_{q\in\mathcal{J}} |S(q)| - |\mathcal{U}|(3\delta L + 4/m)|Y|.
\end{aligned}$$

Substituting this into (81), we get

$$\sum_{q\in\mathcal{J}} |S(q)| \leq \left(\prod_{\mathbf{u}\in\mathcal{U}} (b_{\mathbf{u}} - a_{\mathbf{u}})\right) \cdot |Y| + |\mathcal{U}|(3\delta L + 4/m)|Y| \tag{82}$$

Finally, putting (82) together with (80), we obtain the bound

$$\begin{aligned}\left||\mathcal{S}| - |Y| \cdot \prod_{\mathbf{u}\in\mathcal{U}} (b_{\mathbf{u}} - a_{\mathbf{u}})\right| &\leq |\mathcal{U}|(3\delta L + 4/m) \cdot |Y| \\ &\leq |\mathcal{U}|^2(6L\delta' + 4/m) \cdot |Y|\end{aligned}$$

as required. ■