




On Two-Pass Streaming Algorithms for Maximum Bipartite Matching

Christian Konrad   

Department of Computer Science, University of Bristol, UK

Kheeran K. Naidu   

Department of Computer Science, University of Bristol, UK

Abstract

We study two-pass streaming algorithms for Maximum Bipartite Matching (MBM). All known two-pass streaming algorithms for MBM operate in a similar fashion: They compute a maximal matching in the first pass and find 3-augmenting paths in the second in order to augment the matching found in the first pass. Our aim is to explore the limitations of this approach and to determine whether current techniques can be used to further improve the state-of-the-art algorithms. We give the following results:

We show that every two-pass streaming algorithm that solely computes a maximal matching in the first pass and outputs a $(2/3 + \epsilon)$ -approximation requires $n^{1+\Omega(\frac{1}{\log \log n})}$ space, for every $\epsilon > 0$, where n is the number of vertices of the input graph. This result is obtained by extending the Ruzsa-Szemerédi graph construction of [Goel et al., SODA'12] so as to ensure that the resulting graph has a close to perfect matching, the key property needed in our construction. This result may be of independent interest.

Furthermore, we combine the two main techniques, i.e., subsampling followed by the GREEDY matching algorithm [Konrad, MFCS'18] which gives a $2 - \sqrt{2} \approx 0.5857$ -approximation, and the computation of *degree-bounded semi-matchings* [Esfandiari et al., ICDMW'16][Kale and Tirodkar, APPROX'17] which gives a $\frac{1}{2} + \frac{1}{12} \approx 0.5833$ -approximation, and obtain a meta-algorithm that yields Konrad's and Esfandiari et al.'s algorithms as special cases. This unifies two strands of research. By optimizing parameters, we discover that Konrad's algorithm is optimal for the implied class of algorithms and, perhaps surprisingly, that there is a second optimal algorithm. We show that the analysis of our meta-algorithm is best possible. Our results imply that further improvements, if possible, require new techniques.

2012 ACM Subject Classification Information systems → Data streaming; Mathematics of computing → Matchings and factors; Theory of computation → Communication complexity

Keywords and phrases Data streaming, matchings, lower bounds

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2021.19

Category APPROX

Funding *Kheeran K. Naidu*: EPSRC Doctoral Training Studentship EP/T517872/1.

1 Introduction

In the *semi-streaming model* for processing large graphs, an n -vertex graph is presented to an algorithm as a sequence of its edges in arbitrary order. The algorithm makes one or few passes over the input stream and maintains a memory of size $O(n \text{ polylog } n)$.

The semi-streaming model has been extensively studied since its introduction by Feigenbaum et al. in 2004 [11], and various graph problems, including matchings, independent sets, spanning trees, graph sparsification, subgraph detection, and others are known to admit semi-streaming algorithms (see [23] for an excellent survey). Among these problems, the Maximum Matching problem and, in particular, its bipartite version, the Maximum Bipartite Matching (MBM) problem, have received the most attention (see, for example, [11, 22, 1, 20, 13, 16, 8, 15, 19, 12, 5, 10, 2, 4, 17]).



© Christian Konrad and Kheeran K. Naidu;
licensed under Creative Commons License CC-BY 4.0

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2021).

Editors: Mary Wootters and Laura Sanità; Article No. 19; pp. 19:1–19:18



Leibniz International Proceedings in Informatics

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

Algorithm 1 GREEDY Matching.

Input: Graph $G = (A, B, E)$

- 1: $M \leftarrow \emptyset$
- 2: **for each** edge $e \in E$ (arbitrary order)
- 3: **if** $M \cup \{e\}$ is a matching
- 4: $M \leftarrow M \cup \{e\}$
- 5: **return** M

Algorithm 2 GREEDY _{d} Semi-Matching.

Input: Graph $G = (A, B, E)$, integer d

- 1: $S \leftarrow \emptyset$
- 2: **for each** edge $ab \in E$ (arbitrary order)
- 3: **if** $\deg_S(a) = 0$ **and** $\deg_S(b) < d$
- 4: $S \leftarrow S \cup \{ab\}$
- 5: **return** S

In this paper, we focus on MBM. The currently best one-pass semi-streaming algorithm for MBM is the GREEDY matching algorithm (depicted in Algorithm 1). GREEDY processes the edges of a graph in arbitrary order and inserts the current edge into an initially empty matching if possible. It produces a maximal matching, which is known to be at least half the size of a maximum matching, and constitutes a $\frac{1}{2}$ -approximation semi-streaming algorithm for MBM. It is a long-standing open question whether GREEDY is optimal for the class of semi-streaming algorithms or whether an improved approximation ratio is possible. Progress has been made on the lower bound side ([13, 16, 17]), ruling out semi-streaming algorithms with approximation ratio better than $\frac{1}{1+\ln 2} \approx 0.5906$ [17].

Konrad et al. [20] were the first to show that an approximation ratio better than $\frac{1}{2}$ can be achieved if two passes over the input are allowed, and further successive improvements [15, 8, 19] led to a two-pass semi-streaming algorithm with an approximation factor of $2 - \sqrt{2} \approx 0.58578$ [19] (see Table 1 for an overview of two-pass algorithms for MBM).

Table 1 Two-pass semi-streaming algorithms for Maximum Bipartite Matching.

Approximation Factor	Reference	Comment
$\frac{1}{2} + 0.019$	Konrad et al. [20]	randomized
$\frac{1}{2} + \frac{1}{16} = 0.5625$	Kale and Tirodkar [15]	deterministic
$\frac{1}{2} + \frac{1}{12} \approx 0.5833$	Esfandiari et al. [8]	deterministic
$2 - \sqrt{2} \approx 0.5857$	Konrad [19]	randomized

All known two-pass streaming algorithms proceed in a similar fashion. In the first pass, they run GREEDY in order to compute a maximal matching M . In the second pass, they pursue different strategies to compute additional edges F that allow them to increase the size of M . Two techniques for computing the edge set F have been used:

1. **Subsampling and Greedy** [19] (see also [20]): Given a bipartite graph $G = (A, B, E)$ and a first-pass maximal matching M , they first subsample the edges M with probability p and obtain a matching $M' \subseteq M$. Then, in the second pass, they compute GREEDY matchings M_L and M_R on subgraphs $G_L = G[A(M') \cup \overline{B(M)}]$ and $G_R = G[\overline{A(M)} \cup B(M')]$, respectively, where $A(M')$ are the matched A -vertices in M' , $\overline{B(M)}$ are the unmatched B vertices, and $B(M')$ and $\overline{A(M)}$ are defined similarly. It can be seen that if M is relatively small, then $M' \cup M_L \cup M_R$ contains many disjoint 3-augmenting paths. Setting $p = \sqrt{2} - 1$ yields the approximation factor $2 - \sqrt{2}$.
2. **Semi-matchings and Greedy _{d}** [15, 8]: Given a bipartite graph $G = (A, B, E)$ and a first-pass maximal matching M , the second pass consists of finding *degree- d -constrained semi-matchings* S_L and S_R on subgraphs $G_L = G[A(M) \cup \overline{B(M)}]$ and $G_R = G[\overline{A(M)} \cup B(M)]$, respectively, using the algorithm GREEDY _{d} (as depicted in Algorithm 2). A degree- d -constrained semi-matching in a bipartite graph is a subset of edges $S \subseteq E$ such

that $\deg_S(a) \leq 1$ and $\deg_S(b) \leq d$, for every $a \in A$ and $b \in B$ or vice versa¹. Similar to the method above, it can be seen that if the matching M is relatively small, $M \cup S_L \cup S_R$ contains many disjoint 3-augmenting paths. The setting $d = 3$ yields the approximation factor $\frac{1}{2} + \frac{1}{12}$.

Our Results. In this paper, we explore the limitations of this approach and investigate whether current techniques can be used to further improve the state-of-the-art.

Our first result is a limitation result on the approximation factor achievable by algorithms that follow the scheme described above:

► **Theorem 1 (simplified).** *Every two-pass semi-streaming algorithm for MBM that solely runs GREEDY in the first pass has an approximation factor of at most $\frac{2}{3}$.*

Our result builds upon a result by Goel et al. [13] who proved that the lower bound of Theorem 1 applies to one-pass streaming algorithms. Their construction relies on the existence of dense *Ruzsa-Szemerédi* graphs with large induced matchings, i.e., bipartite $2n$ -vertex graphs $G = (A, B, E)$ with $|A| = |B| = n$ whose edge sets can be partitioned into disjoint induced matchings such that each matching is of size at least $(\frac{1}{2} - \delta)n$, for some small δ . Our construction requires similarly dense RS graphs with equally large matchings, however, in addition to these properties, our RS graphs must contain a *near-perfect* matching, i.e., a matching that matches all but a small constant fraction of the vertices. To this end, we augment the RS graph construction by Goel et al.: We show that, for each induced matching M in Goel et al.'s construction, we can add a matching M' to the construction without violating the induced matching property such that $M \cup M'$ forms a near-perfect matching. We believe this result may be of independent interest.

Next, we combine the subsampling and semi-matching techniques and give a meta-algorithm that yields Konrad's and Esfandiari et al.'s algorithms as special cases, thereby unifying two strands of research. Our meta-algorithm is parameterised by a sampling probability $0 < p \leq 1$ and an integral degree bound $d \geq 1$. First, as in the subsampling technique, the edges of the first-pass matching M are subsampled independently with probability p , which yields a subset $M' \subseteq M$. Next, as in the semi-matching technique, incomplete semi-matchings S_L and S_R with degree bounds d are computed, however, now in the subgraphs $G'_L = G[A(M') \cup \overline{B(M)}]$ and $G'_R = G[\overline{A(M)} \cup B(M')]$. The algorithm then outputs the largest matching among the edges $M \cup S_L \cup S_R$.

As our second result, we establish the approximation factor of our meta-algorithm:

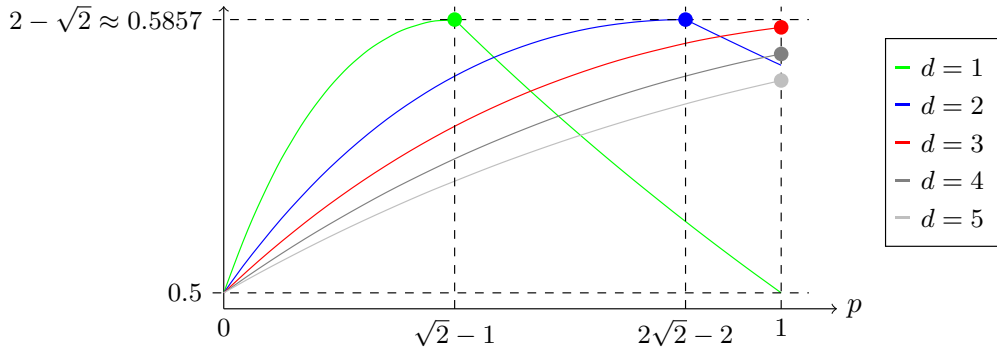
► **Theorem 2 (simplified).** *Combining the subsampling and semi-matching techniques yields a two-pass semi-streaming algorithm for MBM with approximation factor*

$$\begin{cases} \frac{1}{2} + (\frac{1}{d+p} - \frac{1}{2d}) \cdot p, & \text{if } p \leq d(\sqrt{2} - 1) \\ \frac{1}{2} + \frac{d-p}{6d+2p}, & \text{otherwise,} \end{cases}$$

(ignoring lower order terms) that succeeds with high probability.

Interestingly, two parameter settings maximize the approximation factor in Theorem 2, achieving the ratio $2 - \sqrt{2}$ (see Figure 1). This is achieved by setting $d = 1$ and $p = \sqrt{2} - 1$ which recovers Konrad's algorithm, and by setting $d = 2$ and $p = 2\sqrt{2} - 2$ which gives a new algorithm. The setting $d = 3$ and $p = 1$ yields the slightly weaker bound $\frac{1}{2} + \frac{1}{12} \approx 0.5833$ and recovers Esfandiari et al.'s algorithm.

¹ The usual definition of a semi-matching requires $\deg_S(a) = 1$, for every $a \in A$ (e.g. [9, 21]). This property is not required here, and, for ease of notation, we stick to this term.



■ **Figure 1** Approximation factors for different settings of d .

We also show that the analysis of our meta-algorithm is tight, by giving instances on which our meta-algorithm does not perform better than the claimed bound (**Theorem 12**).

Discussion. Our results demonstrate that new techniques are needed in order to improve on the $(2 - \sqrt{2})$ approximation factor. However, one may wonder whether $2 - \sqrt{2}$ is the best approximation ratio achievable by the class of two-pass matching algorithms that solely computes a maximal matching in the first pass. As pointed out by Kapralov [17], his techniques for establishing the $\frac{1}{1+\ln 2}$ lower bound for one-pass algorithms can probably also be applied to a construction by Huang et al. [14], which would then show that $2 - \sqrt{2}$ is the best approximation factor achievable by one-pass semi-streaming algorithms for MBM. It is unclear whether a first-pass GREEDY matching could be embedded in the resulting construction without affecting its hardness, however, if possible, this would render Konrad’s algorithm optimal for the considered class of two-pass streaming algorithms.

Further Related Work. Besides two passes over the input, improvements over the GREEDY algorithm can also be obtained under the assumption that the input stream is in random order. Assadi and Behnezhad [2] recently showed that an approximation factor of $\frac{2}{3} + \epsilon$ can be obtained, for some fixed small but constant $\epsilon > 0$, building on Bernstein’s breakthrough result [5], and improving on previous algorithms [5, 10, 19, 20]. In insertion-deletion streams, where previously inserted edges may be deleted again, space $\tilde{\Theta}(n^{2-3\epsilon})$ is necessary [7] and sufficient [3, 6] for computing a n^ϵ -approximation (see also [18]).

Outline. We first give notation and definitions in Section 2. Subsequently, we show in Section 3 that every two-pass semi-streaming algorithm that solely runs GREEDY in the first pass cannot have an approximation ratio of $\frac{2}{3} + \epsilon$, for any $\epsilon > 0$. Our main algorithmic result, i.e., the combination of subsampling and GREEDY_d , is presented in Section 4. Finally, we conclude in Section 5.

2 Preliminaries

Let $G = (A, B, E)$ be a bipartite graph with $V = A \cup B$ and $|V| = n$. For $F \subseteq E$ and $v \in V$, we write $\text{deg}_F(v)$ to denote the degree of vertex v in subgraph (A, B, F) . For any $U \subseteq V$ and $F \subseteq E$, $U(F)$ denotes the set of vertices in U which are the endpoints of edges in F , and we denote its complement by $\overline{U(F)} = U \setminus U(F)$. For a subset of vertices $U \subseteq V$, we write $G[U]$

for the subgraph of G induced by U . For any edges $e, f \in E$, e is *incident* to f if they share an endpoint. We say that e and f are *vertex-disjoint* if e is not incident to f . Lastly, for any two sets X and Y , we define $X \oplus Y := (X \setminus Y) \cup (Y \setminus X)$ as their symmetric difference.

A *matching* in G is a subset $M \subseteq E$ of vertex-disjoint edges. It is *maximal* if every $e \in E \setminus M$ is incident to an edge in M . We denote by $\mu(G)$ the *matching number* of G , i.e., the cardinality of a largest matching. A *maximum matching* is one of size $\mu(G)$. Additionally, M is called an *induced matching* if the edge set of the subgraph of G induced by $V(M)$ is exactly M .

Wald's Equation. We require the following well-known version of *Wald's Equation*:

► **Lemma 3.** *Let X_1, X_2, \dots be a sequence of non-negative random variables with $\mathbb{E}[X_i] \leq \tau$, for all $i \leq T$, and let T be a random stopping time for the sequence with $\mathbb{E}[T] < \infty$. Then:*

$$\mathbb{E}\left[\sum_{i=1}^T X_i\right] \leq \tau \cdot \mathbb{E}[T] .$$

3 Lower Bound

We now prove that every two-pass streaming algorithm for MBM with approximation factor $\frac{2}{3} + \epsilon$, for any $\epsilon > 0$, that solely runs GREEDY in the first pass requires space $n^{1+\Omega(\frac{1}{\log \log n})}$. To this end, we adapt the lower bound by Goel et al. [13], which we discuss first.

3.1 Goel et al.'s Lower Bound for One-pass Algorithms

Goel et al.'s lower bound is proved in the *one-way two-party communication framework*. Two parties, denoted Alice and Bob, each hold subsets E_1 and E_2 , respectively, of the input graph's edges. Alice sends a single message to Bob who, upon receipt, outputs a large matching. Goel et al. showed that there is a distribution λ over input graphs so that every deterministic communication protocol with constant distributional error over λ and approximation factor $\frac{2}{3} + \epsilon$, for any $\epsilon > 0$, requires a message of length $n^{1+\Omega(\frac{1}{\log \log n})}$. A similar result then applies for randomized constant error protocols by Yao's Lemma [25], and the well-known connection between streaming algorithms and one-way communication protocols allows us to translate this lower bound to a lower bound on the space requirements of constant error one-pass streaming algorithms.

Goel et al.'s construction is based on the existence of a dense Ruzsa-Szemerédi graph:

► **Definition 4** (Ruzsa-Szemerédi Graph). *A bipartite graph $G = (A, B, E)$ is an (r, t) -Ruzsa-Szemerédi graph (RS graph in short) if the edge set E can be partitioned into t disjoint matchings M_1, M_2, \dots, M_t such that, for every i , (1) $|M_i| \geq r$; and (2) M_i is an induced matching in G .*

They give a construction for a family of $((\frac{1}{2} - \delta)n, n^{\Omega(\frac{1}{\log \log n})})$ -RS graphs, for any small constant $\delta > 0$, on $2n$ vertices (with $|A| = |B| = n$) that we will extend further below.

Their hard input distribution λ for the two-party communication setting is displayed in Figure 2. Observe that the graphs $G \sim \lambda$ are such that $\mu(G) \geq \frac{3}{2}N$ since the matching $M_X^* \cup M_Y^* \cup \widehat{M}_s$ is of this size.

Goel et al. prove the following hardness result:

► **Theorem 5.** *For any small $\epsilon > 0$, every deterministic $(\frac{2}{3} + \epsilon)$ -approximation one-way two-party communication protocol with constant distributional error over λ requires a message of size $n^{1+\Omega(\frac{1}{\log \log n})}$, where n is the number of vertices in the input graph.*

1. Let $G^{RS} = (A, B, E)$ be an (r, t) -RS graph with $|A| = |B| = N$ and $r = (\frac{1}{2} - \delta) \cdot N$, for some $\delta > 0$, and $t = N^{\Omega(\frac{1}{\log \log N})}$.
 2. For every $i \in [t]$, let $\widehat{M}_i \subseteq M_i$ be a uniform random subset of size $(\frac{1}{2} - 2\delta) \cdot N$ and let $E_1 = \cup_{i=1}^t \widehat{M}_i$.
 3. Let X and Y each be disjoint sets of $(\frac{1}{2} + \delta) \cdot N$ vertices, which are also disjoint from $A \cup B$. Choose uniformly at random a special index $s \in [t]$.
 4. Let M_X^* and M_Y^* be arbitrary perfect matchings between X and $\overline{B(M_s)}$, and Y and $\overline{A(M_s)}$, respectively. Then, let $E_2 = M_X^* \cup M_Y^*$.
 5. Finally, $G = (A \cup X, B \cup Y, E_1 \cup E_2)$ which has $n = (3 + 2\delta) \cdot N$ vertices.
- Alice is given edges E_1 and Bob is given edges E_2 .

■ Figure 2 Hard input distribution λ .

3.2 Our Lower Bound Construction

In the following, we extend Goel et al.’s lower bound to the two-pass situation where a GREEDY matching is computed in the first pass. To this end, we need to augment Alice and Bob’s inputs, as defined by distribution λ , by a maximal matching M in the input graph $G \sim \lambda$, which then results in a distribution λ^+ . Observe that if we place the edges of M at the beginning of the input stream, then running GREEDY in the first pass recovers exactly the matching M . Hence, when abstracting the second pass as a two-party communication problem, both Alice and Bob already know the matching M . Our main argument then is as follows: We will show that any two-party protocol under distribution λ^+ can also be used for solving the distribution λ with the same distributional error, message size, and similar approximation factor. The hardness of Theorem 5, therefore, carries over.

3.2.1 Ruzsa-Szemerédi Graphs with Near-Perfect Matchings

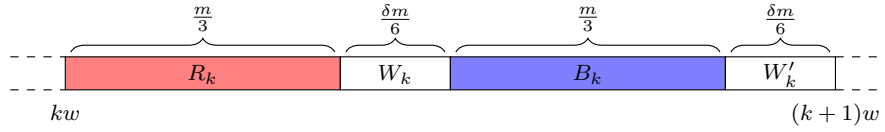
Adding a maximal matching M to Alice’s and Bob’s input requires care since we need to ensure that the hardness of the construction is preserved. Our construction requires that the underlying RS graph contains a near-perfect matching, which is a property that is not guaranteed by Goel et al.’s RS graph construction.

We therefore augment Goel et al.’s construction by complementing every induced matching, M_i , with a vertex-disjoint counterpart, M'_i , without destroying the RS graph properties. Then, since M_i and M'_i are vertex-disjoint, $M_i \cup M'_i$ constitutes a matching, and, since both M_i and M'_i each already match nearly half of the vertices, $M_i \cup M'_i$ constitutes a near-perfect matching in our family of RS graphs.

We will now present Goel et al.’s RS graph construction and then discuss how the additional matchings M'_i can be added to the construction.

Goel et al.’s Ruzsa-Szemerédi Graph Construction

For an integer m , let $X = Y = [m^2]^m$ be the vertex sets of a bipartite graph, and let $N = |X| = |Y| = m^{2m}$ denote their cardinalities. Every induced matching M_I of Goel et al.’s RS graph construction is indexed by a subset of coordinates $I \subset [m]$ of size $\frac{\delta m}{6}$, for some small $\delta > 0$. Then, the edges M_I are defined by means of a colouring of the vertices X and Y (which depends on I), that we discuss first.



■ **Figure 3** One group of the partitioned number line of natural numbers.

Colouring the Vertex Sets. Let $w = \frac{(2+\delta)m}{3}$. Then, define a partition of the natural numbers into groups of size w such that, for all $k \in \mathbb{N}_0$,

$$\begin{aligned}
 R_k &= \left[kw, kw + \frac{m}{3} \right) && \text{where } |R_k| = \frac{m}{3}, \\
 W_k &= \left[kw + \frac{m}{3}, kw + \frac{m}{3} + \frac{\delta m}{6} \right) && \text{where } |W_k| = \frac{\delta m}{6}, \\
 B_k &= \left[kw + \frac{m}{3} + \frac{\delta m}{6}, kw + \frac{2m}{3} + \frac{\delta m}{6} \right) && \text{where } |B_k| = \frac{m}{3}, \\
 W'_k &= \left[kw + \frac{2m}{3} + \frac{\delta m}{6}, (k+1)w \right) && \text{where } |W'_k| = \frac{\delta m}{6}.
 \end{aligned}$$

See Figure 3 for an illustration.

Given I , let $L_s = \{\vec{x} \in [m^2]^m : \sum_{i \in I} x_i = s\}$ represent a layer of vectors in $[m^2]^m$ where the 1-norm of their subvectors² w.r.t. I is s , for all $s \in \mathbb{N}_0$. Next, colour the vectors in each L_s either red if $s \in R_k$, blue if $s \in B_k$, or white if $s \in W_k \cup W'_k$, for some $k \in \mathbb{N}_0$. Doing this gives the following coloured strips for any $k \in \mathbb{N}_0$ (see Figure 4):

$$R(k) = \bigcup_{\forall s \in R_k} L_s, \quad W(k) = \bigcup_{\forall s \in W_k} L_s, \quad B(k) = \bigcup_{\forall s \in B_k} L_s \quad \text{and} \quad W'(k) = \bigcup_{\forall s \in W'_k} L_s.$$

Next, these strips are grouped together by colour, as follows:

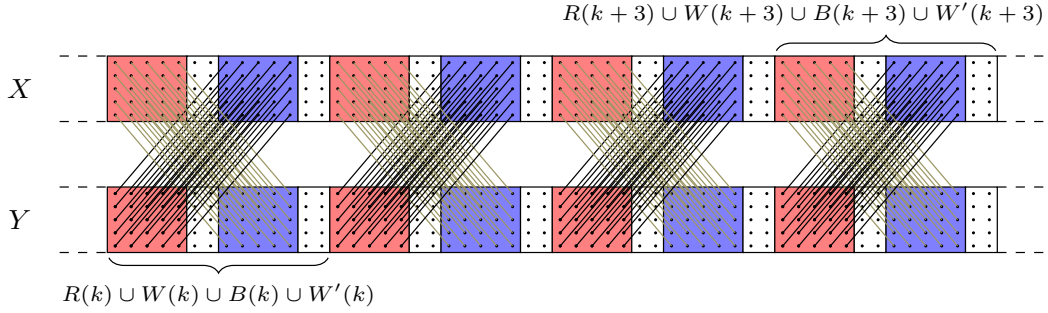
$$R = \bigcup_{\forall k \in \mathbb{N}_0} R(k), \quad W = \bigcup_{\forall k \in \mathbb{N}_0} W(k), \quad B = \bigcup_{\forall k \in \mathbb{N}_0} B(k) \quad \text{and} \quad W' = \bigcup_{\forall k \in \mathbb{N}_0} W'(k).$$

We now define the colours of the vertices X and Y as follows: A vertex $\vec{z} \in X \cup Y$ is coloured red if $\vec{z} \in R$, blue if $\vec{z} \in B$, and white if $\vec{z} \in W \cup W'$. Let $R^X = R \cap X$ and define $B^X, W^X, W'^X, R^Y, B^Y, W^Y, W'^Y$ similarly.

Definition of the Induced Matchings. Goel et al. construct the edges of the induced matching M_I by pairing every blue vertex $\vec{b} \in B^X$ with each coordinate greater than $\frac{2}{\delta} + 1$ to a red vertex $\vec{r} \in R^Y$, such that $\vec{r} = \vec{b} - (\frac{2}{\delta} + 1) \cdot \vec{1}_I$, where $\vec{1}_I$ is the characteristic vector of set I . See Figure 4 for an illustration.

Goel et al. show that M_I is large, i.e., $|M_I| \geq (\frac{1}{2} - \delta) \cdot N - o(N)$. Observe that any two distinct indexing sets I and J produce their own vertex colourings and matchings M_I and M_J . They prove that, as long as the index sets I and J have a sufficiently small intersection (at most $(\frac{5\delta}{12})(\frac{\delta m}{6})$), M_I and M_J are induced matchings w.r.t. to each other. Hence, they show the existence of a large family \mathcal{T} , with $|\mathcal{T}| = N^{\Omega(\frac{1}{\log \log N})}$, of subsets $I \subset [m]$ whose pairwise intersections are of size at most $(\frac{5\delta}{12})(\frac{\delta m}{6})$. Then, the matchings of the RS graph are identified as the matchings M_I , for every $I \in \mathcal{T}$.

² A subvector in this context is the result of a trivial mapping of the vector to a lower dimensional subspace.



■ **Figure 4** Illustration of the vertex colouring and induced matchings for a fixed I . The black edges are M_I and the gold ones are M'_I .

Extending Goel et al.'s Construction

For every indexing set $I \in \mathcal{T}$ and respective matching M_I of Goel et al.'s construction, we symmetrically construct an additional matching M'_I by pairing every blue vertex in Y (instead of X), $\vec{b} \in B^Y$, with each coordinate greater than $\frac{2}{\delta} + 1$, to a red vertex in X , $\vec{r} \in R^X$, such that $\vec{r} = \vec{b} - (\frac{2}{\delta} + 1) \cdot \vec{1}_I$. See Figure 4 for an illustration.

We immediately see that, by virtue of being symmetrical, $|M'_I| = |M_I| (\geq (\frac{1}{2} - \delta) \cdot N - o(N))$.

Furthermore, by construction, M'_I and M_I are vertex-disjoint matchings, hence $M_I \cup M'_I$ is a matching, and, taking their respective sizes into account, $M_I \cup M'_I$ is a near-perfect matching as required. Since, for any distinct $I, J \in \mathcal{T}$, M_I and M_J are induced matchings w.r.t. each other, the symmetrical nature of our additional matchings implies the same for M'_I and M'_J . However, showing that M_I and M'_J are induced with respect to each other is not immediately clear. Fortunately, Goel et al.'s proof already implicitly shows this, and, for completeness, we reproduce the decisive argument:

► **Lemma 6.** *Given two distinct sets of indices I and J such that $|I \cap J| \leq (\frac{5\delta}{12})(\frac{\delta m}{6})$, no edge in M_I is induced by M'_J , for any small enough $\delta > 0$.*

Proof. Let $\vec{b} \in B^X$ be matched to $\vec{r} \in R^Y$ by M_I , i.e., $\vec{b} - \vec{r} = (\frac{2}{\delta} + 1) \cdot \vec{1}_I$. If the edge (\vec{b}, \vec{r}) is induced by M'_J , then one endpoint is coloured blue and the other red in the colouring of X and Y with respect to J . Hence, \vec{b} and \vec{r} are separated by a single white strip (see Figure 4) and

$$|\sum_{j \in J} (\vec{b} - \vec{r})_j| \geq \frac{\delta m}{6}. \quad (1)$$

On the other hand,

$$|\sum_{j \in J} (\vec{b} - \vec{r})_j| = |\sum_{j \in J} ((\frac{2}{\delta} + 1) \cdot \vec{1}_I)_j| = (\frac{2}{\delta} + 1) \cdot |I \cap J| \leq (\frac{5}{6} + \frac{5\delta}{12})(\frac{\delta m}{6}),$$

which contradicts Equation 1 for small enough δ . ◀

We thus obtain the following theorem:

► **Theorem 7.** *For any small enough constant $\delta > 0$, there exists a family of bipartite (r, t) -Ruzsa-Szemerédi graphs where $|A| = |B| = N$, $r = (\frac{1}{2} - \delta) \cdot N$, and $t = N^{\Omega(\frac{1}{\log \log N})}$ such that there are $N^{\Omega(\frac{1}{\log \log N})}$ disjoint near-perfect matchings each of size exactly $(1 - 2\delta) \cdot N$.*

1. Let G^{RS} be an RS graph as in Theorem 7. Fix some induced matching M_i and let $M_i \cup M'_i$ be its near-perfect matching of size $(1 - 2\delta) \cdot N$.
 2. Let F be an arbitrary set of $2\delta N$ additional edges such that $P = M_i \cup M'_i \cup F$ is a perfect matching in G^{RS} .
 3. Consider distribution λ constructed using RS graph $G^{RS} \setminus (M_i \cup M'_i)$.
 4. For every $G = (V, E) \sim \lambda$, let $P_G = M_i \cup M'_i \cup (F \setminus E)$ (to avoid multi-edges) and add P_G to G to obtain the input graph G^+ .
- The edges $P \cup E_1$ are given to Alice and the edges $P \cup E_2$ are given to Bob (recall that E_1 and E_2 are defined in distribution λ).

■ **Figure 5** Hard input distribution λ^+ .

3.2.2 Lower Bound Proof

Equipped with RS graphs with near-perfect matchings and input distribution λ , we now define our hard input distribution λ^+ , see Figure 5.

We are now ready to prove our main lower bound theorem:

► **Theorem 8.** *For any $\epsilon > 0$, every deterministic $(\frac{2}{3} + \epsilon)$ -approximation one-way communication protocol with constant distributional error over λ^+ for MBM requires a message of size $n^{1+\Omega(\frac{1}{\log \log n})}$, where n is the number of vertices in the input graph.*

Proof. Let γ^+ be a deterministic $(\frac{2}{3} + \epsilon)$ -approximation protocol that solves distribution λ^+ with constant distributional error. Given γ^+ , we will now define a protocol γ that solves distribution λ with the same communication cost, same error, and approximation ratio strictly better than $\frac{2}{3}$. Invoking Theorem 5 then proves our result.

The protocol γ is easy to obtain: Observe that P in distribution λ^+ is the same for every sampled input graph $G^+ \sim \lambda^+$. Hence, in protocol γ , Alice and Bob first make sure that the edges P are included in their inputs. This is achieved by Alice adding the edges $P \setminus E_1 = P_G$ to her input, and Bob adding the edges P to his input. In doing so, Alice and Bob's input is equivalently distributed to choosing an input graph G^+ from λ^+ . Alice and Bob can, therefore, run protocol γ^+ which produces an output matching M_{out}^+ . Bob then outputs the largest matching M_{out} among the edges $M_X^* \cup M_Y^* \cup (M_{\text{out}}^+ \setminus P_G)$ as the output of the protocol γ .

Next, we will argue that $|M_{\text{out}}| \geq |M_{\text{out}}^+| - |F| = |M_{\text{out}}^+| - 2\delta N$. We can construct a matching \tilde{M} of this size as follows: First, add every edge $e \in M_{\text{out}}^+$ that is not contained in P to \tilde{M} . Second, for every edge $e \in M_{\text{out}}^+ \cap (M_i \cup M'_i)$, we insert the incident edge to e that is contained in $M_X^* \cup M_Y^*$ into \tilde{M} (notice that these incident edges always exist except for edges from the special induced matching). This implies that $|M_{\text{out}}| \geq |\tilde{M}| \geq |M_{\text{out}}^+| - |F|$.

Recall that $\mu(G) \geq \frac{2}{3}N$ and, since G is a subgraph of G^+ , $\mu(G^+) \geq \mu(G)$. This implies that $N \leq \frac{2}{3}\mu(G^+)$. Since γ^+ is a $(\frac{2}{3} + \epsilon)$ -approximation protocol, we have $|M_{\text{out}}^+| \geq (\frac{2}{3} + \epsilon)\mu(G^+)$, and thus:

$$|M_{\text{out}}| \geq |M_{\text{out}}^+| - 2\delta N \geq (\frac{2}{3} + \epsilon)\mu(G^+) - 2\delta \frac{2}{3}\mu(G^+) = (\frac{2}{3} + \epsilon - \frac{4}{3}\delta)\mu(G^+).$$

Hence, setting $\delta < \frac{3}{4}\epsilon$ in distribution λ yields a protocol with approximation ratio strictly above $\frac{2}{3}$. This, however, implies that γ requires a message of length $n^{1+\Omega(\frac{1}{\log \log n})}$ (Theorem 5), and since the message sent in γ and γ^+ is equivalent, the result follows. ◀

Applying Yao's Lemma and the usual connection between streaming algorithms and one-way communication protocols, we obtain our main lower bound result:

■ **Algorithm 3** Finding Augmenting Paths.

Input: A stream of edges π of a bipartite graph $G = (A, B, E)$, a maximal matching M in G , $p \in (0, 1]$ and $d \in \mathbb{N}^+$.

- 1: Let $M' \subseteq M$ be a random subset such that $\forall e \in M, \Pr[e \in M'] = p$
- 2: Let $G'_L = G[A(M') \cup \overline{B(M)}]$ and $G'_R = G[\overline{A(M)} \cup B(M')]$
- 3: Denote by $\pi_{G'_L}$ ($\pi_{G'_R}$) the substream of π of edges of G'_L (G'_R , respectively)
- 4: $S_L \leftarrow \text{GREEDY}_d(\pi_{G'_L})$ such that $\deg_{S_L}(a) \leq 1$, for every $a \in A(M')$, and $\deg_{S_L}(b) \leq d$, for every $b \in \overline{B(M)}$
- 5: $S_R \leftarrow \text{GREEDY}_d(\pi_{G'_R})$ such that $\deg_{S_R}(b) \leq 1$, for every $b \in B(M')$, and $\deg_{S_R}(a) \leq d$, for every $a \in \overline{A(M)}$
- 6: $\mathcal{P} \leftarrow \{ab', ab, a'b : ab' \in S_L, ab \in M', a'b \in S_R\}$
- 7: **return** A largest subset $\mathcal{Q} \subseteq \mathcal{P}$ of vertex-disjoint paths.

► **Theorem 1.** For any $\epsilon > 0$, every (possibly randomised) two-pass streaming algorithm for MBM with approximation ratio $\frac{2}{3} + \epsilon$ that solely computes a GREEDY matching in the first pass requires $n^{1+\Omega(\frac{1}{\log \log n})}$ space, where n is the number of vertices in the graph.

4 Algorithm

In this section, we combine the subsampling approach as used by Konrad [19] and the semi-matching approach as used by Esfandiari et al. [8] and Kale and Tirodkar [15] in order to find many disjoint 3-augmenting paths, see Algorithm 3.

The input to Algorithm 3 is a stream of edges π of a bipartite graph $G = (A, B, E)$, a maximal matching M in G (e.g., computed in a first pass by GREEDY), a sampling probability p , and an integral degree bound d . First, each edge of M is included in M' with probability p . Then, while processing the stream, degree- d -bounded semi-matchings S_L and S_R are computed using the algorithm GREEDY_d (see Algorithm 2 in Section 1). The algorithm then returns a largest subset of vertex-disjoint 3-augmenting paths \mathcal{Q} . We can thus obtain a matching of size $|M| + |\mathcal{Q}|$.

4.1 Analysis of Algorithm 3

The main task in analysing Algorithm 3 is to bound the sizes of S_L and S_R from below. A bound that holds in expectation for the case $d = 1$ was previously proved by Konrad et al. [20], and a high probability result (for $d = 1$) was later obtained by Konrad [19]. We also first give a bound that holds in expectation (Lemma 9), which is achieved by extending the original proof by Konrad et al. [20]. Our extension, however, is non-trivial as it requires a very different progress measure. Then, following Konrad [19], we obtain a high probability version in Lemma 10.

We also remark that Lemmas 9 and 10 are stated in a more general context, however, it is not hard to see that they capture the situation of the computations of S_L and S_R in subgraphs G'_L and G'_R , respectively.

► **Lemma 9.** Let $G = (A, B, E)$ be a bipartite graph, π an arbitrarily ordered stream of its edges, $p \in (0, 1]$, and $d \in \mathbb{N}^+$. Let $A' \subseteq A$ be a random subset such that $\forall a \in A, \Pr[a \in A'] = p$, and let d be the degree bound of the B vertices. Let $H = G[A' \cup B]$ and denote by π_H the substream of π consisting of the edges in H . Then,

$$\mathbb{E}_{A'}[|\text{GREEDY}_d(\pi_H)|] \geq \frac{d}{d+p} \cdot p \cdot \mu(G).$$

Proof. Let M^* be a fixed maximum matching in G and let $M_H^* := \{ab \in M^* : a \in A'\}$ be the subset of edges incident to A' .

Game Setup. Consider the following game: On selection of an edge by $\text{GREEDY}_d(\pi_H)$, the edge *attacks* the (at most two) incident edges of M_H^* and deals damage to them. Initially, the damage of every edge in M_H^* is 0, and the maximum damage of each such edge is 1. A damage below 1 means that the edge could still be selected by the algorithm. A damage equal to 1 implies that the edge can no longer be selected.

Denote by S_i the first i edges selected by $\text{GREEDY}_d(\pi_H)$ and let ab be the $(i+1)^{\text{th}}$ edge selected. The way damage is dealt is as follows:

- If there is an edge $a'b \in M_H^*$ such that $a' \notin A(S_{i+1})$ then attack edge $a'b$ by adding $\frac{1}{d}$ damage to it;
- If there is an edge $ab' \in M_H^*$ then attack edge ab' by adding $1 - \frac{\deg_{S_i}(b')}{d}$ damage to it, maxing out the damage to 1.

Observe that the maximum damage an edge selected by $\text{GREEDY}_d(\pi_H)$ can inflict is at most $1 + \frac{1}{d}$ (applying both cases to the two incident optimal edges). Furthermore, observe that the maximum damage every edge in M_H^* receives is 1, and, indeed, at the end of the algorithm, every edge in M_H^* has damage 1.

Applying Wald's Equation. Denote by s the cardinality of the semi-matching computed by $\text{GREEDY}_d(\pi_H)$ and let X_1, X_2, \dots, X_s be the sequence of edges selected. Define the random variable Y_i to be the damage dealt by edge X_i . Let T be the smallest i such that $\sum_{j=1}^i Y_j = |M_H^*|$ holds. Observe that T is a random stopping time. To apply the version of Wald's Equation presented in Lemma 3, we need to show that $\mathbb{E}[T]$ is finite and find a value τ such that, for all $i \leq T$, $\mathbb{E}[Y_i] \leq \tau$ holds:

The expected stopping time $\mathbb{E}[T]$ is finite since $T \leq s$ always holds by the end of the algorithm, i.e., the total damage dealt is $|M_H^*|$. Finding τ is less obvious. By definition, the damage Y_i dealt by any edge X_i is either 0, $\frac{1}{d}, \dots, 1$ or $1 + \frac{1}{d}$. Hence, we obtain the following:

$$\mathbb{E}[Y_i] \leq \Pr[Y_i \leq 1] \cdot 1 + \underbrace{\Pr\left[Y_i = 1 + \frac{1}{d}\right]}_q \cdot \left(1 + \frac{1}{d}\right) = (1 - q) \cdot 1 + q \cdot \left(1 + \frac{1}{d}\right) = 1 + \frac{q}{d}.$$

It remains to bound $\Pr[Y_i = 1 + \frac{1}{d}] (= q)$. Let $X_i = ab$. Then, by definition of the game, the event $Y_i = 1 + \frac{1}{d}$ only happens if there exists an edge $a'b \in M_H^*$ such that $a' \notin A(S_i)$. In this case, ab inflicts a damage of 1 on edge $a'b$. However, observe that since $a' \notin A(S_i)$, the random choice as to whether $a' \in A'$ and thus whether $a'b \in M_H^*$ had not needed to occur yet (principle of deferred decision). Hence, we obtain:

$$\Pr[Y_i = 1 + \frac{1}{d}] \leq \Pr[a' \in A'] = p.$$

Having shown that $\mathbb{E}[T]$ is finite and $\mathbb{E}[Y_i] \leq 1 + \frac{p}{d}$ for all $i \leq T$, we can apply Wald's Equation (Lemma 3) and we obtain $\mathbb{E}[\sum_{j=1}^T Y_j] \leq (1 + \frac{p}{d})\mathbb{E}[T]$. Finally, since $\mathbb{E}[\sum_{j=1}^T Y_j] = \mathbb{E}[|M_H^*|] = p \cdot \mu(G)$ and $T \leq s = |\text{GREEDY}_d(\pi_H)|$, it follows that

$$\mathbb{E}\left[\sum_{j=1}^T Y_j\right] = p \cdot \mu(G) \leq \left(1 + \frac{p}{d}\right) \cdot \mathbb{E}[T] \leq \left(1 + \frac{p}{d}\right) \cdot \mathbb{E}[|\text{GREEDY}_d(\pi_H)|],$$

which implies the result. ◀

19:12 On Two-Pass Streaming Algorithms for Maximum Bipartite Matching

Next, we follow the approach by Konrad [19] to strengthen Lemma 9 and obtain the following high probability result (see Appendix A for the proof):

► **Lemma 10.** *Let $G = (A, B, E)$ be a bipartite graph, π be any arbitrary stream of its edges, $p \in (0, 1]$ and $d \in \mathbb{N}^+$. Let $A' \subseteq A$ be a random subset such that $\forall a \in A, \Pr[a \in A'] = p$, let d be the degree bound of the B vertices and let $H = G[A' \cup B]$. Then, the following holds with probability at least $1 - 2\mu(G)^{-18}$:*

$$|\text{GREEDY}_d(\pi_H)| \geq \frac{d}{d+p} \cdot p \cdot \mu(G) - o(\mu(G)).$$

Equipped with Lemma 10, we are now ready to bound the number of augmenting paths found by Algorithm 3.

► **Lemma 11.** *Suppose that $|M| = (\frac{1}{2} + \epsilon)\mu(G)$. Then, with probability $1 - \mu(G)^{-16}$, the number of vertex-disjoint 3-augmenting paths $|\mathcal{Q}|$ found by Algorithm 3 is at least:*

$$|\mathcal{Q}| \geq \left(\frac{1-2\epsilon}{d+p} - \frac{1+2\epsilon}{2d}\right) \cdot p \cdot \mu(G) - o(\mu(G)).$$

Proof. Let M^* be a fixed maximum matching in G . In this proof, we will refer to the quantities used by Algorithm 3. First, using a Chernoff bound for independent Poisson trials, we see that $|M'| = p \cdot |M| \pm O(\sqrt{|M| \ln |M|})$ with probability at least $1 - |M|^{-C}$ for an arbitrarily large constant C .

Consider the subgraphs $G_L = G[A(M) \cup \overline{B(M)}]$ and $G_R = G[\overline{A(M)} \cup B(M)]$. $M \oplus M^*$ contains $(\frac{1}{2} - \epsilon)\mu(G)$ vertex-disjoint augmenting paths where each path starts and ends with an edge in $G_L \cup G_R$. This implies that

$$\mu(G_L) + \mu(G_R) \geq 2\left(\frac{1}{2} - \epsilon\right)\mu(G) = (1 - 2\epsilon)\mu(G). \quad (2)$$

Following Konrad [19], we will argue next that

$$|\mathcal{P}| \geq |S_L| + |S_R| - |M'|. \quad (3)$$

Observe that there are $|M'| - |S_L|$ vertices of $|M'|$ that are not incident to an edge in S_L , and similarly, $|M'| - |S_R|$ vertices of $|M'|$ that are not incident to an edge in S_R . Hence, there are at least $|M'| - (|M'| - |S_L|) - (|M'| - |S_R|) = |S_L| + |S_R| - |M'|$ edges of $|M'|$ that are incident to both an edge from S_L and S_R . We thus obtain that there are at least $|\mathcal{P}| \geq |S_L| + |S_R| - |M'|$ 3-augmenting paths.

Next, Esfandiari et al. (Lemma 6 in [8]) consider a similar structure to \mathcal{P} and argue that there is at least a d -fraction of augmenting paths in \mathcal{P} that are vertex-disjoint, and, hence,

$$|\mathcal{Q}| \geq \frac{1}{d}|\mathcal{P}|. \quad (4)$$

Using Lemma 10 and Inequalities 2, 3, and 4, we obtain:

$$\begin{aligned} |\mathcal{Q}| &\geq \frac{1}{d}(|S_L| + |S_R| - |M'|) \\ &\geq \frac{1}{d} \left(\frac{d}{d+p} \cdot p \cdot (1 - 2\epsilon)\mu(G) - o(\mu(G)) - p \cdot \left(\frac{1}{2} + \epsilon\right)\mu(G) - O(\sqrt{\mu(G) \ln \mu(G)}) \right) \\ &= \left(\frac{1-2\epsilon}{d+p} - \frac{1+2\epsilon}{2d}\right) \cdot p \cdot \mu(G) - o(\mu(G)). \end{aligned}$$

Using the union bound, the error of the algorithm is bounded by $|M|^{-C} + 2\mu(G)^{-18} \leq \mu(G)^{-16}$. ◀

1. Let $A_{\text{in}} = \{a_{\text{in}}^1, a_{\text{in}}^2, \dots, a_{\text{in}}^N\}$, $A_{\text{out}} = \{a_{\text{out}}^1, \dots, a_{\text{out}}^N\}$, $B_{\text{in}} = \{b_{\text{in}}^1, \dots, b_{\text{in}}^N\}$, and $B_{\text{out}} = \{b_{\text{out}}^1, \dots, b_{\text{out}}^N\}$ be sets of vertices, for some integer N .
2. Let $M = \{a_{\text{in}}^i b_{\text{in}}^i : 1 \leq i \leq N\}$ be a perfect matching between A_{in} and B_{in} . Let $G_L = (A_{\text{in}}, B_{\text{out}}, E_L)$ be a semi-complete graph such that $a_{\text{in}}^i b_{\text{out}}^j \in E_L \Leftrightarrow i \geq j$, and let $G_R = (A_{\text{out}}, B_{\text{in}}, E_R)$ be a semi-complete graph such that $a_{\text{out}}^i b_{\text{in}}^j \in E_R \Leftrightarrow i \geq j$.
3. Our bipartite hard instance graph is defined as $G = (A_{\text{in}} \cup A_{\text{out}}, B_{\text{in}} \cup B_{\text{out}}, M \cup E_L \cup E_R)$ and has $n = 4N$ vertices.
4. Finally, let π be a stream of its edges where the edges of M arrive first followed by the edges E_L and E_R . The edges in E_L are ordered so that $a_{\text{in}}^i b_{\text{out}}^j$ arrives before $a_{\text{in}}^{i'} b_{\text{out}}^{j'}$ only if $i > i'$, or $i = i'$ and $j < j'$. Similarly, the edges in E_R are ordered so that $a_{\text{out}}^i b_{\text{in}}^j$ arrives before $a_{\text{out}}^{i'} b_{\text{in}}^{j'}$ only if $i > i'$, or $i = i'$ and $j < j'$.

■ **Figure 6** Hard input instance G for Algorithm 3.

We are now ready to state our main algorithmic result:

► **Theorem 2.** *For every $p \in (0, 1]$ and every integral $d \geq 1$, there is a two-pass semi-streaming algorithm for MBM with approximation factor*

$$\begin{cases} \frac{1}{2} + \left(\frac{1}{d+p} - \frac{1}{2d}\right) \cdot p - o(1), & \text{if } p \leq d(\sqrt{2} - 1) \\ \frac{1}{2} + \frac{d-p}{6d+2p} - o(1), & \text{otherwise,} \end{cases}$$

that succeeds with high probability (in $\mu(G)$, where G is the input graph). The settings ($d = 1, p = \sqrt{2} - 1$) and ($d = 2, p = 2(\sqrt{2} - 1)$) maximize the approximation factor to $2 - \sqrt{2} - o(1)$.

Proof. Let M be a maximal matching such that $|M| = (\frac{1}{2} + \epsilon)\mu(G)$, for some $0 \leq \epsilon \leq \frac{1}{2}$ and some bipartite graph $G = (A, B, E)$ with a stream π of its edges. Let \mathcal{Q} be the disjoint augmenting paths found by Algorithm 3 on input π, M, p and d . Then, augmenting M with \mathcal{Q} yields a matching of size $|M| + |\mathcal{Q}|$. By Lemma 11, the following inequality holds with high probability:

$$|M| + |\mathcal{Q}| \geq \left(\frac{1}{2} + \epsilon\right)\mu(G) + \left(\frac{1-2\epsilon}{d+p} - \frac{1+2\epsilon}{2d}\right) \cdot p \cdot \mu(G) - o(\mu(G)). \quad (5)$$

We distinguish two cases:

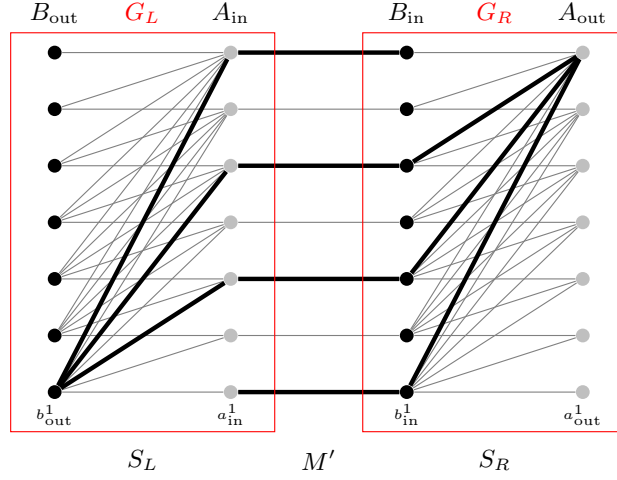
1. If $p \leq d(\sqrt{2} - 1)$ then $\epsilon = 0$ minimizes the RHS of Inequality 5, and we obtain the claimed bound by plugging the value $\epsilon = 0$ into the inequality.
2. If $p \geq d(\sqrt{2} - 1)$ (only possible if $d \in \{1, 2\}$) then $\epsilon = \frac{d-p}{6d+2p}$ minimizes the RHS of Inequality 5, and we obtain the claimed bound by plugging the value $\epsilon = \frac{d-p}{6d+2p}$ into the inequality.

It can be seen that, for a fixed d , the maximum is obtained if $p = \min\{d\sqrt{2} - d, 1\}$, and the values $d \in \{1, 2\}$ yield the claimed bound of $2 - \sqrt{2} - o(1)$ (see Figure 1 in Section 1). ◀

4.2 Optimality of the Analysis

We will show now that our analysis of Algorithm 3 is best possible. To this end, we define a worst-case input graph G in Figure 6, and prove in Theorem 12 that Algorithm 3 does not perform better on G than predicted by our analysis. See Figure 7 for an illustration.

Observe that M is a maximal matching in G , and if we run GREEDY in the first pass on π then M would be returned. Let $M_L^* = \{a_{\text{in}}^i b_{\text{out}}^i : 1 \leq i \leq N\}$ and $M_R^* = \{a_{\text{out}}^i b_{\text{in}}^i : 1 \leq i \leq N\}$. Then, M_L^* is a perfect matching in G_L , M_R^* is a perfect matching in G_R , and $M_L^* \cup M_R^*$ is a perfect matching in G .



■ **Figure 7** Algorithm 3 on a hard input instance with $N = 7$, $d = 3$ and $p = 0.5$.

► **Theorem 12.** *Algorithm 3 with parameters $d \geq 1$ and $0 < p \leq 1$ on input G received via stream π and maximal matching M finds at most*

$$\left(\left(\frac{1}{d+p} - \frac{1}{2d} \right) \cdot p + o(1) \right) \mu(G)$$

augmenting paths with high probability. This renders our analysis of Algorithm 3 best possible when $p \leq d\sqrt{2} - d$.

Proof. In this proof, we will refer to the quantities used by Algorithm 3, that is, M' (the edges of M subsampled with probability p), S_L and S_R .

We will use the following claim in our proof:

▷ **Claim 13.** With high probability, for every pair $i, j \in [N]$ with $i \leq j$, we have

$$|\{a_{\text{in}}^k b_{\text{in}}^k \in M' \mid i \leq k \leq j\}| = p \cdot (j - i) \pm o(N).$$

Proof. This claim is easy to prove. Indeed, for any fixed $i, j \in [N]$ with $i \leq j$, the statement above follows directly from the Chernoff bound. Using the union bound over all pairs $i, j \in [N]$, the claim follows. ◁

From now on, we condition on the event that the statement in Claim 13 holds.

Let $A'_{\text{in}} = A(M')$ and let $B'_{\text{in}} = B(M')$. We will first argue that, for two different vertices $a_{\text{in}}^i, a_{\text{in}}^j \in A'_{\text{in}}$ with $i < j$, if $a_{\text{in}}^i \in A(S_L)$ then $a_{\text{in}}^j \in A(S_L)$ also holds. Indeed, suppose that this was not the case. Let b_{out}^k be the partner of a_{in}^i in S_L . Observe that the edges $a_{\text{in}}^i b_{\text{out}}^k, a_{\text{in}}^j b_{\text{out}}^k \in E_L$, and, in particular, the edge $a_{\text{in}}^j b_{\text{out}}^k$ arrives before the edge $a_{\text{in}}^i b_{\text{out}}^k$ in π . Hence, edge $a_{\text{in}}^j b_{\text{out}}^k$ would have been selected, a contradiction. A similar argument holds for vertices $b_{\text{out}}^i, b_{\text{out}}^j \in B_{\text{out}}$ with $i > j$; if $\deg_{S_L}(b_{\text{out}}^i) \geq 1$ then $\deg_{S_L}(b_{\text{out}}^j) = d$.

Let i_{min} be the smallest index such that $a_{\text{in}}^{i_{\text{min}}} \in A(S_L)$. We will now argue that $i_{\text{min}} \geq \frac{pN}{p+d} - o(N)$. Observe that the vertices A'_{in} are matched in order from the largest to smallest index, and each matched vertex in A'_{in} is matched only once. The vertices in B_{out} are matched from the smallest to largest index, and each matched vertex is matched d times (except possibly the last such matched vertex). Consider the last edge $a_{\text{in}}^{i_{\text{min}}} b_{\text{out}}^q$ inserted into S_L . Then, $q \leq i_{\text{min}}$, and, thus, $|B(S_L)| \leq i_{\text{min}}$. By Claim 13 (applied with $j = N$), we have

$|A(S_L)| \geq p \cdot (N - i_{\min}) - o(N)$ with high probability. Since $|A(S_L)|$ is matched to $B(S_L)$ in S_L , and each B -vertex is matched at most d times, we obtain $|A(S_L)| \leq d \cdot |B(S_L)|$, and, hence:

$$p \cdot (N - i_{\min}) - o(N) \leq |A(S_L)| \leq d \cdot |B(S_L)| \leq d \cdot i_{\min} ,$$

which implies $i_{\min} \geq \frac{pN}{p+d} - o(N)$.

Let i_{\max} be the largest index such that $b_{\text{in}}^{i_{\max}} \in B(S_R)$. Using a similar argument as above, we see that $i_{\max} \leq \frac{dN}{p+d} + o(N)$.

Let $M'' = \{a_{\text{in}}^i b_{\text{in}}^i \in M' : i_{\min} \leq i \leq i_{\max}\}$ be the subset of augmentable edges, i.e., edges for which there exists a left wing in S_L and a right wing in S_R . Then, by Claim 13, we have

$$|M''| \leq p \cdot (i_{\max} - i_{\min}) + o(N) \leq \frac{p(d-p)N}{p+d} + o(N) .$$

All but constantly many vertices in $A(M'')$ share the same neighbour in S_L with $d-1$ other vertices of $A(M'')$. Hence, at most a d -fraction (plus up to the constantly many exceptions, which disappear in the $o(N)$ term) of M'' can be augmented simultaneously. Using $N = \frac{1}{2}\mu(G)$, we obtain the following bound on the number of edges that can be augmented simultaneously:

$$\frac{1}{d}|M''| \leq \frac{1}{d} \left(\frac{p(d-p)N}{p+d} + o(N) \right) = \left(\left(\frac{1}{d+p} - \frac{1}{2d} \right) \cdot p + o(1) \right) \mu(G) . \quad \blacktriangleleft$$

5 Conclusion

In this paper, we studied the class of two-pass semi-streaming algorithms for MBM that solely compute a GREEDY matching in the first pass. We showed that algorithms of this class cannot have an approximation ratio of $\frac{2}{3} + \epsilon$, for any $\epsilon > 0$. We also combined the two dominant techniques that have previously been used for designing such algorithms and discovered another algorithm that matches the state-of-the-art approximation factor of $2 - \sqrt{2} \approx 0.58578$.

We conclude with two open problems. First, we are particularly interested in whether there exists a one-pass semi-streaming algorithm that is able to augment a maximal matching so as to yield an approximation ratio above $2 - \sqrt{2}$. Second, is there a two-pass semi-streaming algorithm for MBM that improves on the approximation factor of $2 - \sqrt{2}$ and operates differently in the first pass to the class of algorithms considered in this paper?

References

- 1 Kook Jin Ahn and Sudipto Guha. Linear programming in the semi-streaming model with application to the maximum matching problem. In *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. doi:10.1007/978-3-642-22012-8_42.
- 2 Sepehr Assadi and Soheil Behnezhad. Beating two-thirds for random-order streaming matching. In Nikhil Bansal, Emanuela Merelli, and James Worrell, editors, *48th International Colloquium on Automata, Languages, and Programming, ICALP 2021, July 12-16, 2021, Glasgow, Scotland (Virtual Conference)*, volume 198 of *LIPICs*, pages 19:1–19:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPICs.ICALP.2021.19.

19:16 On Two-Pass Streaming Algorithms for Maximum Bipartite Matching

- 3 Sepehr Assadi, Sanjeev Khanna, Yang Li, and Grigory Yaroslavtsev. Maximum matchings in dynamic graph streams and the simultaneous communication model. In Robert Krauthgamer, editor, *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 1345–1364. SIAM, 2016. doi:10.1137/1.9781611974331.ch93.
- 4 Sepehr Assadi, S. Cliff Liu, and Robert E. Tarjan. An auction algorithm for bipartite matching in streaming and massively parallel computation models. In Hung Viet Le and Valerie King, editors, *4th Symposium on Simplicity in Algorithms, SOSA 2021, Virtual Conference, January 11-12, 2021*, pages 165–171. SIAM, 2021. doi:10.1137/1.9781611976496.18.
- 5 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. doi:10.4230/LIPICs.ICALP.2020.12.
- 6 Rajesh Chitnis, Graham Cormode, Hossein Esfandiari, MohammadTaghi Hajiaghayi, Andrew McGregor, Morteza Monemizadeh, and Sofya Vorotnikova. Kernelization via sampling with applications to finding matchings and related problems in dynamic graph streams. In Robert Krauthgamer, editor, *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 1326–1344. SIAM, 2016. doi:10.1137/1.9781611974331.ch92.
- 7 Jacques Dark and Christian Konrad. Optimal lower bounds for matching and vertex cover in dynamic graph streams. In Shubhangi Saraf, editor, *35th Computational Complexity Conference, CCC 2020, July 28-31, 2020, Saarbrücken, Germany (Virtual Conference)*, volume 169 of *LIPICs*, pages 30:1–30:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPICs.CCC.2020.30.
- 8 Hossein Esfandiari, MohammadTaghi Hajiaghayi, and Morteza Monemizadeh. Finding large matchings in semi-streaming. In Carlotta Domeniconi, Francesco Gullo, Francesco Bonchi, Josep Domingo-Ferrer, Ricardo Baeza-Yates, Zhi-Hua Zhou, and Xindong Wu, editors, *IEEE International Conference on Data Mining Workshops, ICDM Workshops 2016, December 12-15, 2016, Barcelona, Spain*, pages 608–614. IEEE Computer Society, 2016. doi:10.1109/ICDMW.2016.0092.
- 9 Jittat Fakcharoenphol, Bundit Laekhanukit, and Danupon Nanongkai. Faster algorithms for semi-matching problems. *ACM Trans. Algorithms*, 10(3), 2014. doi:10.1145/2601071.
- 10 Alireza Farhadi, Mohammad Taghi Hajiaghayi, Tung Mai, Anup Rao, and Ryan A. Rossi. Approximate maximum matching in random streams. In Shuchi Chawla, editor, *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pages 1773–1785. SIAM, 2020. doi:10.1137/1.9781611975994.108.
- 11 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. doi:10.1007/978-3-540-27836-8_46.
- 12 Buddhima Gamlath, Sagar Kale, Slobodan Mitrovic, and Ola Svensson. Weighted matchings via unweighted augmentations. In Peter Robinson and Faith Ellen, editors, *Proceedings of the 2019 ACM Symposium on Principles of Distributed Computing, PODC 2019, Toronto, ON, Canada, July 29 - August 2, 2019*, pages 491–500. ACM, 2019. doi:10.1145/3293611.3331603.
- 13 Ashish Goel, Michael Kapralov, and Sanjeev Khanna. On the communication and streaming complexity of maximum bipartite matching. In Yuval Rabani, editor, *Proceedings of the 23rd ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages pp. 468–485. SIAM, 2012. doi:10.1137/1.9781611973099.41.

- 14 Zhiyi Huang, Binghui Peng, Zhihao Gavin Tang, Runzhou Tao, Xiaowei Wu, and Yuhao Zhang. Tight competitive ratios of classic matching algorithms in the fully online model. 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 2875–2886. SIAM, 2019. doi:10.1137/1.9781611975482.178.
- 15 Sagar Kale and Sumedh Tirodkar. Maximum matching in two, three, and a few more passes over graph streams. In Klaus Jansen, José D. P. Rolim, David Williamson, and Santosh S. Vempala, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2017, August 16-18, 2017, Berkeley, CA, USA*, volume 81 of *LIPICs*, pages 15:1–15:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017. doi:10.4230/LIPICs.APPROX-RANDOM.2017.15.
- 16 Michael Kapralov. Better bounds for matchings in the streaming model. In Sanjeev Khanna, editor, *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 1679–1697. SIAM, 2013. doi:10.1137/1.9781611973105.121.
- 17 Michael Kapralov. Space lower bounds for approximating maximum matching in the edge arrival model. In Dániel Marx, editor, *Proceedings of the 32nd ACM-SIAM Symposium on Discrete Algorithms, (SODA)*, pages pp. 1874–1893. SIAM, 2021. doi:10.1137/1.9781611976465.112.
- 18 Christian Konrad. Maximum matching in turnstile streams. In Nikhil Bansal and Irene Finocchi, editors, *Algorithms - ESA 2015 - 23rd Annual European Symposium, Patras, Greece, September 14-16, 2015, Proceedings*, volume 9294 of *Lecture Notes in Computer Science*, pages 840–852. Springer, 2015. doi:10.1007/978-3-662-48350-3_70.
- 19 Christian Konrad. A simple augmentation method for matchings with applications to streaming algorithms. In Igor Potapov, Paul G. Spirakis, and James Worrell, editors, *43rd International Symposium on Mathematical Foundations of Computer Science, MFCS 2018, August 27-31, 2018, Liverpool, UK*, volume 117 of *LIPICs*, pages 74:1–74:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPICs.MFCS.2018.74.
- 20 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. doi:10.1007/978-3-642-32512-0_20.
- 21 Christian Konrad and Adi Rosén. Approximating semi-matchings in streaming and in two-party communication. *ACM Trans. Algorithms*, 12(3), 2016. doi:10.1145/2898960.
- 22 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 International Workshop 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. doi:10.1007/11538462_15.
- 23 Andrew McGregor. Graph stream algorithms: a survey. *SIGMOD Rec.*, 43(1):9–20, 2014. doi:10.1145/2627692.2627694.
- 24 Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005. doi:10.1017/CB09780511813603.
- 25 Andrew Chi-Chih Yao. Probabilistic computations: Toward a unified measure of complexity. In *Proceedings of the 18th Symposium on Foundations of Computer Science (FOCS)*, pages pp. 222–227. IEEE Computer Society, 1977. doi:10.1109/SFCS.1977.24.

A

 Strengthening Lemma 9

Following [19], we use tail inequalities for martingales to strengthen Lemma 9 and give a high probability result. The proof of Lemma 10 uses the *Azuma-Hoeffding's Inequality* [24, Theorem 12.4]:

► **Lemma 14** (Azuma-Hoeffding's Inequality). *Let Z_0, Z_1, \dots, Z_n be a martingale such that $\forall k \geq 0, |Z_{k+1} - Z_k| \leq c_k$. Then, $\forall t \geq 0$ and any $\lambda > 0$,*

$$\Pr[|Z_t - Z_0| \geq \lambda] \leq 2 \exp\left(\frac{-\lambda^2}{2 \sum_{k=0}^{t-1} c_k^2}\right).$$

► **Lemma 10.** *Let $G = (A, B, E)$ be a bipartite graph, π be any arbitrary stream of its edges, $p \in (0, 1]$ and $d \in \mathbb{N}^+$. Let $A' \subseteq A$ be a random subset such that $\forall a \in A, \Pr[a \in A'] = p$, let d be the degree bound of the B vertices and let $H = G[A' \cup B]$. Then, the following holds with probability at least $1 - 2\mu(G)^{-18}$:*

$$|\text{GREEDY}_d(\pi_H)| \geq \frac{d}{d+p} \cdot p \cdot \mu(G) - o(\mu(G)).$$

Proof. Let X_1, X_2, \dots, X_s be the sequence of random variables representing the edges selected by $\text{GREEDY}_d(\pi_H)$ with the source of randomness from the choice of A' . Define $Y := |\text{GREEDY}_d(\pi_H)|$. Then, we define the random variables $Z_i := \mathbb{E}[Y | X_1, \dots, X_i]$ for all $i = 0, \dots, s$ to be the corresponding Doob Martingale, and let $Z_i = Z_{i-1}$, for every $i > s$. Notice that $Z_s = Y$ and $Z_0 = \mathbb{E}[Y] \geq \frac{d}{d+p} \cdot p \cdot \mu(G)$ by Lemma 9. Now, we will show that any deviation of Y from its expectation, $|Z_s - Z_0|$, is small with high probability.

To that end, we first need to bound $|Z_{i+1} - Z_i|$ for all $i \geq 0$. Notice that $|Z_{i+1} - Z_i| = 0$ for all $i \geq s$. Next, we will argue that $|Z_{i+1} - Z_i| \leq 1$ for all $i < s$. Indeed, for any fixed first i edges added to the semi-matching, any two different choices for X_{i+1} yield two potentially different semi-matchings S_1, S_2 , respectively, such that $S_1 \oplus S_2$ consists of at most one alternating path. Hence, the two semi-matchings differ by at most one edge, which proves the claim.

Then, we have that $s = Y \leq d \cdot \mu(H) \leq d \cdot \mu(G)$ and it follows that $|Z_{i+1} - Z_i| \leq 1$ for all $i \leq d \cdot \mu(G)$ and $|Z_{i+1} - Z_i| = 0$ for all $i > d \cdot \mu(G)$. Finally, by applying Azuma-Hoeffding's Inequality (see Lemma 14), we finalise the proof:

$$\Pr\left[|Z_s - Z_0| \geq 6\sqrt{d\mu(G) \ln \mu(G)}\right] \leq 2\mu(G)^{-18},$$

where $|Z_s - Z_0| = |Y - \mathbb{E}[Y]|$. ◀