On the communication and streaming complexity of maximum bipartite matching

Ashish Goel* Michael Kapralov† Sanjeev Khanna‡

November 27, 2011

Abstract

Consider the following communication problem. Alice holds a graph $G_A = (P, Q, E_A)$ and Bob holds a graph $G_B = (P, Q, E_B)$, where $|P| = |Q| = n$. Alice is allowed to send Bob a message m that depends only on the graph G_A. Bob must then output a matching $M \subseteq E_A \cup E_B$. What is the minimum message size of the message m that Alice sends to Bob that allows Bob to recover a matching of size at least $(1 - \epsilon)$ times the maximum matching in $G_A \cup G_B$? The minimum message length is the *one-round communication complexity* of approximating bipartite matching. It is easy to see that the one-round communication complexity also gives a lower bound on the space needed by a one-pass streaming algorithm to compute a $(1 - \epsilon)$-approximate bipartite matching. The focus of this work is to understand one-round communication complexity and one-pass streaming complexity of maximum bipartite matching. In particular, how well can one approximate these problems with linear communication and space? Prior to our work, only a $\frac{1}{2}$-approximation was known for both these problems.

In order to study these questions, we introduce the concept of an ϵ-matching cover of a bipartite graph G, which is a sparse subgraph of the original graph that preserves the size of maximum matching between every subset of vertices to within an additive ϵn error. We give a polynomial time construction of a $\frac{1}{2}$-matching cover of size $O(n)$ with some crucial additional properties, thereby showing that Alice and Bob can achieve a $\frac{2}{3}$-approximation with a message of size $O(n)$. While we do not provide bounds on the size of ϵ-matching covers for $\epsilon < 1/2$, we prove that in general, the size of the smallest ϵ-matching cover of a graph G on n vertices is essentially equal to the size of the largest so-called ϵ-Ruzsa Szemerédi graph on n vertices. We use this connection to show that for any $\delta > 0$, a $(\frac{2}{3} + \delta)$-approximation requires a communication complexity of $n^{1+\Omega(1/\log\log n)}$.

We also consider the natural restrictingon of the problem in which G_A and G_B are only allowed to share vertices on one side of the bipartition, which is motivated by applications to one-pass streaming with vertex arrivals. We show that a $\frac{3}{4}$-approximation can be achieved with a linear size message in this case, and this result is best possible in that super-linear space is needed to achieve any better approximation.

Finally, we build on our techniques for the restricted version above to design one-pass streaming algorithm for the case when vertices on one side are known in advance, and the vertices on the other side arrive in a streaming manner together with all their incident edges. This is precisely the setting of the celebrated $(1 - \frac{1}{e})$-competitive randomized algorithm of Karp-Vazirani-Vazirani (KVV) for the *online* bipartite matching problem [12]. We present here the first *deterministic* one-pass streaming $(1 - \frac{1}{e})$-approximation algorithm using $O(n)$ space for this setting.

1 Introduction

We study the communication and streaming complexity of the maximum bipartite matching problem. Consider the following scenario. Alice holds a graph $G_A = (P, Q, E_A)$ and Bob holds a graph $G_B = (P, Q, E_B)$, where $|P| = |Q| = n$. Alice is allowed to send Bob a message m that depends only on the graph G_A. Bob must then output a matching $M \subseteq E_A \cup E_B$. What is the minimum size of the message m that Alice sends to Bob that allows Bob to recover a matching of size at least $1 - \epsilon$ of the maximum matching in $G_A \cup G_B$? The minimum message length is the *one-round communication complexity* of approximating bipartite matching, and is denoted by $CC(\epsilon, n)$. It is easy to see that the quantity $CC(\epsilon, n)$ also gives a lower bound on the space needed by a one-pass streaming algorithm to compute a $(1 - \epsilon)$-approximate bipartite matching. To see this, consider the graph $G_A \cup G_B$ revealed in a streaming manner with edge set E_A revealed first (in some arbitrary order), followed by the edge set E_B. It is clear that any non-trivial approximation to the bipartite matching problem requires $\Omega(n)$ communication and $\Omega(n)$ space, respectively, for the one–round communication and one-

*Departments of Management Science and Engineering and (by courtesy) Computer Science, Stanford University. Email: ashishg@stanford.edu. Research supported in part by NSF award IIS-0904325.

†Institute for Computational and Mathematical Engineering, Stanford University. Email: kapralov@stanford.edu. Research supported in part by NSF award IIS-0904325 and a Stanford Graduate Fellowship.

‡Department of Computer and Information Science, University of Pennsylvania, Philadelphia PA. Email: sanjeev@cis.upenn.edu. Supported in part by NSF Awards CCF-1116961 and IIS-0904314.

pass streaming problems described above. The central question considered in this work is how well can we approximate the bipartite matching problem when only $\tilde{O}(n)$ communication/space is allowed.

Matching Covers: We show that a study of these questions is intimately connected to existence of sparse "matching covers" for bipartite graphs. An *ϵ-matching cover* or simply an ϵ-cover, of a graph $G(P, Q, E)$ is a subgraph $G'(P, Q, E')$ such that for any pairs of sets $A \subseteq P$ and $B \subseteq Q$, the graph G' preserves the size of the largest A to B matching to within an additive error of ϵn. The notion of matching sparsifiers may be viewed as a natural analog of the notion of cut-preserving sparsifiers which have played a very important role in the study of network design and connectivity problems [11, 4]. It is easy to see that if there exists an ϵ-cover of size $f(\epsilon, n)$ for some function f, then Alice can just send a message of size $f(\epsilon, n)$ to allow Bob to compute an additive ϵn error approximation to bipartite matching (and $(1-\epsilon)$-approximation whenever $G_A \cup G_B$ contains a perfect matching). However, we show that the question of constructing efficient ϵ-covers is essentially equivalent to resolving a long-standing problem on a family of graphs known as the *Ruzsa-Szemerédi graphs*. A bipartite graph $G(P, Q, E)$ is an *ϵ-Ruzsa-Szemerédi graph* if E can be partitioned into a collection of induced matchings of size at least ϵn each. Ruzsa-Szemerédi graphs have been extensively studied as they arise naturally in property testing, PCP constructions and additive combinatorics [7, 10, 17]. A major open problem is to determine the maximum number of edges possible in an ϵ-Ruzsa-Szemerédi graph. In particular, do there exist dense graphs with large locally sparse regions (i.e. large induced subgraphs are perfect matchings)? We establish the following somewhat surprising relationship between matching covers and Ruzsa-Szemerédi graphs: for any $\epsilon > 0$ the smallest possible size of an ϵ-matching cover is essentially equal to the largest possible number of edges in an ϵ-Ruzsa-Szemerédi graph.

Constructing dense ϵ-Ruzsa-Szemerédi graphs for general ϵ and proving upper bounds on their size appears to be a difficult problem [9]. To our knowledge, there are two known constructions in the literature. The original construction due to Ruzsa and Szemerédi yields a collection of $n/3$ induced matchings of size $n/2^{O(\sqrt{\log n})}$ using Behrend's construction of a large subset of $\{1, \ldots, n\}$ without three-term arithmetic progressions [3, 17]. Constructions of a collection of $n^{c/\log\log n}$ induced matchings of size $n/3 - o(n)$ were given in [7, 15]. We use the ideas of [7, 15] to construct $(\frac{1}{2} - \delta)$-Ruzsa-Szemerédi graphs with $n^{1+\Omega_\delta(1/\log\log n)}$ edges and a more general construction for the vertex arrival case. To the best of our knowledge, the only known upper bound on the size of ϵ-Ruzsa-Szemerédi graphs for constant $\epsilon < \frac{1}{2}$ is $O(n^2/\log^* n)$ that follows from the bound used in an elementary proof of Roth's theorem [17].

One-round Communication: We show that in fact $CC(\epsilon, n) \leq 2n - 1$ for all $\epsilon \geq \frac{1}{3}$, i.e. a message of linear size suffices to get a $\frac{2}{3}$-approximation to the maximum matching in $G_A \cup G_B$. We establish this result by constructing an $O(n)$ size $\frac{1}{2}$-cover of the input graph that satisfies certain additional properties which allows Bob to recover a $\frac{2}{3}$-approximation[1]. We refer to this particular $\frac{1}{2}$-cover as a *matching skeleton* of the input graph, and give a polynomial time algorithm for constructing it. Next, building on the above-mentioned connection between matching covers and Ruzsa-Szemerédi graphs, we show the following two results: (a) our construction of $\frac{1}{2}$-cover implies that for any $\delta > 0$, there do not exist $(\frac{1}{2} + \delta)$-Ruzsa-Szemerédi graph with more than $O(n/\delta)$ edges, and (b) our $\frac{2}{3}$-approximation result is best possible when only linear amount of communication is allowed. In particular, Alice needs to send $n^{1+\Omega(1/\log\log n)}$ bits to achieve a $(\frac{2}{3}+\delta)$-approximation, for any constant $\delta > 0$, even when randomization is allowed.

We then study the one round communication complexity $CC_v(\epsilon, n)$ of $(1-\epsilon)$-approximate maximum matching in the restricted model when the graphs G_A and G_B are only allowed to share vertices on one side of the bipartition. This model is motivated by application to one-pass streaming computations when the vertices of the graph arrive together with all incident edges. We obtain a stronger approximation result in this model, namely, using the preceding $\frac{1}{2}$-cover construction we show that $CC_v(\epsilon, n) \leq 2n - 1$ for $\epsilon \geq 1/4$. Thus a $\frac{3}{4}$-approximation can be obtained with linear communication complexity, and as before, we show that obtaining a better approximation requires a communication complexity of $n^{1+\Omega(1/\log\log n)}$ bits.

One-pass Streaming: We build on our techniques for one-round communication to design a one-pass streaming algorithm for the case when vertices on one side are known in advance, and the vertices on the other side arrive in a streaming manner together with all their incident edges. This is precisely the setting of the celebrated $(1 - \frac{1}{e})$-competitive randomized algorithm of Karp-Vazirani-Vazirani (KVV) for the *online* bipartite matching problem [12]. We give a *deterministic* one-pass streaming algorithm that matches the $(1 - \frac{1}{e})$-approximation guarantee of KVV using only

[1] We note here that a maximum matching in a graph is only a $\frac{2}{3}$-cover.

$O(n)$ space. Prior to our work, the only known *deterministic* algorithm for matching in one-pass streaming model, even under the assumption that vertices arrive together with all their edges, is the trivial algorithm that keeps a maximal matching, achieving a factor of $\frac{1}{2}$. We note that in the online setting, randomization is crucial as no deterministic online algorithm can achieve a competitive ratio better than $\frac{1}{2}$.

Related work: The streaming complexity of maximum bipartite matching has received significant attention recently. Space-efficient algorithms for approximating maximum matchings to factor $(1-\epsilon)$ in a number of passes that only depends on $1/\epsilon$ have been developed. The work of [14] gave the first space-efficient algorithm for finding matchings in general (non-bipartite) graphs that required a number of passes dependent only on $1/\epsilon$, although the dependence was exponential. This dependence was improved to polynomial in [5], where $(1-\epsilon)$-approximation was obtain in $O(1/\epsilon^8)$ passes. In a recent work, [2] obtained a significant improvement, achieving $(1-\epsilon)$-approximation in $O(\log\log(1/\epsilon)/\epsilon^2)$ passes (their techniques also yield improvements for the weighted version of the problem). Further improvements for the non-bipartite version of the problem have been obtained in [1]. Despite the large body of work on the problem, the only known algorithm for one pass is the trivial algorithm that keeps a maximal matching. No non-trivial lower bounds on the space complexity of obtaining constant factor approximation to maximum bipartite matching in one pass were known prior to our work (for exact computation, an $\Omega(n^2)$ lower bound was shown in [6]).

Organization: We start by introducing relevant definitions in section 2. In section 3 we give the construction of the *matching skeleton*, which we use later in section 4 to prove that $CC(1/3, n) = O(n)$, as well as show that the matching skeleton forms a 1/2-cover. In section 5 we deduce using the matching skeleton that $CC_v(1/4, n) = O(n)$. In section 6 we use these techniques to obtain a deterministic one-pass $(1-1/e)$ approximation to maximum matching in $O(n)$ space in the vertex arrival model. We extend the construction of Ruzsa-Szemerédi graphs from [7, 15] in section 7. We use these extensions in section 8 to show that our upper bounds on $CC(\epsilon, n)$ and $CC_v(\epsilon, n)$ are best possible, as well as to prove lower bounds on the space complexity of one-pass algorithms for approximating maximum bipartite matching. Finally, in section 9 we prove the correspondence between the size of the smallest ϵ-matching cover of a graph on n nodes and the size of the largest ϵ-Ruzsa-Szemerédi graph on n nodes.

2 Preliminaries

We start by defining bipartite matching covers, which are matchings-preserving graph sparsifiers.

DEFINITION 2.1. *Given an undirected bipartite graph $G = (P, Q, E)$, and sets $A \subseteq P, B \subseteq Q$, and $H \subseteq E$, let $M_H(A, B)$ denote the size of the largest matching in the graph $G' = (A, B, (A \times B) \cap H)$.*

Given an undirected bipartite graph $G = (P, Q, E)$ with $|P| = |Q| = n$, a set of edges $H \subseteq E$ is said to be an *ϵ-matching-cover* of G if for all $A \subseteq P, B \subseteq Q$, we have $M_H(A, B) \geq M_E(A, B) - \epsilon n$.

DEFINITION 2.2. *Define $L_C(\epsilon, n)$ to be the smallest number m' such that any undirected bipartite graph $G = (P, Q, E)$ with $P = Q = n$ has an ϵ-matching-cover of size at most m'.*

We next define induced matchings and Ruzsa-Szemerédi graphs.

DEFINITION 2.3. *Given an undirected bipartite graph $G = (P, Q, E)$ and a set of edges $F \subseteq E$, let $P(F) \subseteq P$ denote the set of vertices in P which are incident on at least one edge in F, and analogously, let $Q(F)$ denote the set of vertices in Q which are incident on at least one edge in F. Let $E(F)$, called the set of edges induced by F, denote the set of edges $E \cap (P(F) \times Q(F))$. Note that $E(F)$ may be much larger than F in general.*

Given an undirected bipartite graph $G = (P, Q, E)$, a set of edges $F \subseteq E$ is said to be an *induced matching* if no two edges in F share an endpoint, and $E(F) = F$. Given an undirected bipartite graph $G = (P, Q, E)$ and a partition $\mathcal{F}$ of E, the partition is said to be an *induced partition* of G if every set $F \in \mathcal{F}$ is an induced matching. An undirected bipartite graph $G = (P, Q, E)$ with $P = Q = n$ is said to have an *ϵ-induced partition* if there exists an induced partition of G such every set in the partition is of size at least ϵn. Following [7], we refer to graphs that have an ϵ-induced partition as *ϵ-Ruzsa-Szemerédi graphs.*

DEFINITION 2.4. *Let $U_I(\epsilon, n)$ denote the largest number m such that there exists an undirected bipartite graph $G = (P, Q, E)$ with $|E| = m, |P| = |Q| = n$, and with an ϵ-induced partition.*

Note that for any $0 < \epsilon_1 < \epsilon_2 < 1$, any ϵ_2-induced partition of a graph is also an ϵ_1-induced partition, and hence, $U_I(\epsilon, n)$ is a non-increasing function of ϵ. Analogously, any ϵ_1-matching-cover is also an ϵ_2-matching cover, and hence, $L_C(\epsilon, n)$ is also a non-increasing function of ϵ.

3 Matching Skeletons

Let $G = (P, Q, E)$ be a bipartite graph. We now define a subgraph $G' = (P, Q, E')$ of G that contains at most $(|P| + |Q| - 1)$ edges, and encodes useful information about matchings in G. We refer to this subgraph G' as a *matching skeleton* of G, and this construction will serve as a building block for our algorithms. Among other things, we will show later that G' is a $\frac{1}{2}$-cover of G.

We present the construction of G' in two steps. We first consider the case when P is *hypermatchable*, that is, for every vertex $v \in Q$ there exists a perfect matching of the P side that does not include v. We then extend the construction to the general case using the Edmonds-Gallai decomposition [16].

3.1 P is hypermatchable in G We note that since P is *hypermatchable*, by Hall's theorem [16], we have that $|\Gamma(A)| > |A|$ for all $A \subseteq P$. For a parameter $\alpha \in (0, 1]$, let $\mathcal{R}_G(\alpha) = \{A \subseteq P : |\Gamma_G(A)| \leq (1/\alpha)|A|\}$. Note that as the parameter α *decreases*, the expansion requirement in the definition above *increases*. We will omit the subscript G when G is fixed, as in the next lemma.

LEMMA 3.1. *Let $\alpha \in (0, 1]$ be such that $\mathcal{R}(\alpha+\epsilon) = \emptyset$ for any $\epsilon > 0$, i.e. G supports an $\frac{1}{\alpha+\epsilon}$-matching of the P-side for any $\epsilon > 0$. Then for any two $A_1 \in \mathcal{R}(\alpha), A_2 \in \mathcal{R}(\alpha)$ one has $A_1 \cup A_2 \in \mathcal{R}(\alpha)$.*

Proof. Let $B_1 = \Gamma(A_1)$ and $B_2 = \Gamma(A_2)$. First, since $(A_1 \times (Q \setminus B_1)) \cap E = \emptyset$ and $(A_2 \times (Q \setminus B_2)) \cap E = \emptyset$, we have that $(A_1 \cap A_2) \times (Q \setminus (B_1 \cap B_2)) = \emptyset$. Furthermore, since $\mathcal{R}(\alpha + \epsilon) = \emptyset$, one has $|B_1 \cap B_2| \geq (1/\alpha)|A_1 \cap A_2|$. Also, we have $|B_i| \leq |A_i|/\alpha, i = 1, 2$. Hence,

$$|B_1 \cup B_2| = |B_1| + |B_2| - |B_1 \cap B_2|$$
$$\leq (1/\alpha)(|A_1| + |A_2| - |A_1 \cap A_2|) = (1/\alpha)|A_1 \cup A_2|,$$

and thus $(A_1 \cup A_2) \in \mathcal{R}(\alpha)$ as required.

We now define a collection of sets $(S_j, T_j), j = 1, \ldots, +\infty$, where $S_j \subseteq P, T_j \subseteq Q, S_i \cap S_j = \emptyset, i \neq j$.

1. Set $j := 1$, $G_0 := G, \alpha_0 := 1$. We have $\mathcal{R}_{G_0}(\alpha_0) = \emptyset$.
2. Let $\beta < \alpha_{j-1}$ be the largest real such that $\mathcal{R}_{G_{j-1}}(\beta) \neq \emptyset$.
3. Let $S_\beta = \bigcup_{A \in \mathcal{R}(\beta)} A$, and $T_\beta = \Gamma(S_\beta)$. We have $S_\beta \in \mathcal{R}_{R_{j-1}}(\beta)$ by Lemma 3.1.
4. Let $G_j := G_{j-1} \setminus (S_\beta \cup T_\beta)$. We refer to the value of α at which a pair (S_α, T_α) gets removed from the graph as the expansion of the pair. Set $S_j := S_\beta, T_j := T_\beta, \alpha_j := \beta$. If $G_j \neq \emptyset$, let $j := j+1$ and go to (2).

The following lemma is an easy consequence of the above construction.

LEMMA 3.2. *1. For each $U \subseteq S_j$ one has $|\Gamma_{G_j}(U)| \geq (1/\alpha_j)|U|$.*

2. For every $k > 0$,

$$\left(\left(\bigcup_{j \leq k} S_j\right) \times \left(Q \setminus \bigcup_{j \leq k} T_j\right)\right) \cap E = \emptyset.$$

Proof. We prove (1) by contradiction. When $j = 1$, (1) follows immediately since we are choosing the largest β such that $\mathcal{R}(\beta) \neq \emptyset$. Otherwise suppose that there exists $U \subseteq P_{G_j}$ such that $|\Gamma_{G_j}(U)| < (1/\alpha_j)|U|$. Then first observe that $|\Gamma_{G_j}(U)| > (1/\alpha_{j-1})|U|$. If not then

$$|\Gamma_{G_{j-1}}(S_{j-1} \cup U)| = |T_{j-1}| + |\Gamma_{G_j}(U)|$$
$$\leq \frac{1}{\alpha_{j-1}}(|S_{j-1}| + |U|) \leq \frac{1}{\alpha_{j-1}}(|S_{j-1} \cup U|),$$

since $S_{j-1} \cap P_{G_j} = \emptyset$ by construction. Now as $\alpha_j < \alpha_{j-1}$ is chosen to be the largest real for which there exists some subset $U' \subseteq P_{G_j}$ with $|\Gamma_{G_j}(U')| \leq (1/\alpha_j)|U'|$, it follows that for every $U \subseteq P_{G_j}$, we must have $|\Gamma_{G_j}(U)| \geq (1/\alpha_j)|U|$.

(2) follows by construction.

To complete the definition of the matching skeleton, we now identify the set of edges of G that our algorithm keeps. For a parameter $\gamma \geq 1$ and subsets $S \subseteq P$, $T \subseteq Q$ we refer to a (fractional) matching M that saturates each vertex in S exactly γ times (fractionally) and each vertex in T at most once as a *γ-matching* of S in $(S, T, (S \times T) \cap E)$. By Lemma 3.2 there exists a (fractional) $(1/\alpha_j)$-matching of S_j in $(S_j, T_j, (S_j \times T_j) \cap E)$. Moreover, one can ensure that the matching is supported on the edges of a forest by rerouting flow along cycles. Let M_j be a fractional $(1/\alpha_j)$-matching in (S_j, T_j) that is a forest.

Interestingly, the fractional matching corresponding to the matching skeleton is identical to a 1-majorized fractional allocation of unit-sized jobs to $(1 - \infty)$ machines [13, 8]; as a result, the fractional matchings x_e simultaneously minimize all convex functions of the x_e's subject to the constraint that every node in P is matched exactly once.

3.2 General bipartite graphs We now extend the construction to general bipartite graphs using the Edmonds-Gallai decomposition of $G(P, Q, E)$, which essentially allows us to partition the vertices of G into sets $A_P(G)$, $D_P(G)$, $C_P(G)$, $A_Q(G)$, $D_Q(G)$, and $C_Q(G)$ such that $A_P(G)$ is hypermatchable to $D_Q(G)$, $A_Q(G)$ is hypermatchable to $D_P(G)$, and there is a perfect matching between $C_P(G)$ and $C_Q(G)$.

The Edmonds-Gallai decomposition theorem is as follows.

THEOREM 3.1. *(Edmonds-Gallai decomposition, [16]) Let $G = (V, E)$ be a graph. Then V can be partitioned into the union of sets $D(G), A(G), C(G)$ such that*

$$D(G) = \{v \in V | \exists \text{ a maximum matching missing } v\}$$
$$A(G) = \Gamma(D(G))$$
$$C(G) = V \setminus (D(G) \cup A(G)).$$

Moreover, every maximum matching contains a perfect matching inside $C(G)$.

Applying Edmonds-Gallai decomposition to bipartite graphs, we get

COROLLARY 3.1. *Let $G = (P, Q, E)$ be a graph. Then V can be partitioned into the union of sets $D_P(G), D_Q(G)$, $A_P(G), A_Q(G)$, $C_P(G), C_Q(G)$ such that*

$$D_P(G) = \{v \in P | \exists \text{ a maximum matching missing } v\}$$
$$D_Q(G) = \{v \in Q | \exists \text{ a maximum matching missing } v\}$$
$$A_P(G) = \Gamma(D_Q(G))$$
$$A_Q(G) = \Gamma(D_P(G))$$
$$C_P(G) = P \setminus (D_P(G) \cup A_P(G))$$
$$C_Q(G) = Q \setminus (D_Q(G) \cup A_Q(G)).$$

Moreover,

1. *there exists a perfect matching between $C_P(G)$ and $C_Q(G)$*
2. *for every $U \subseteq A_P(G)$ one has $|\Gamma(U) \cap D_Q(G)| > |U|$*
3. *for every $U \subseteq A_Q(G)$ one has $|\Gamma(U) \cap D_P(G)| > |U|$.*

Proof. (1) is part of the statement of Theorem 3.1. To show (2), note that by definition of $D_Q(G)$ for each vertex $v \in D_Q(G)$ there exists a maximum matching that misses v. Thus, $|\Gamma(U) \cap D_Q(G)| > |U|$ for every set U.

Using the above partition, we can now define a matching skeleton of G. Let $S_0 = C_P(G), T_0 = C_Q(G)$, and let M_0 be a perfect matching between S_0 and T_0. Let $(S_1, T_1), \ldots, (S_j, T_j)$ be the expanding pairs obtained by the construction in the previous section on the graph induced by $A_P(G) \cup D_Q(G)$. Let $(S_{-j}, T_{-j}), \ldots, (S_{-1}, T_{-1})$ be the expanding pairs obtained by the construction in the previous section from the Q side on the graph induced by $A_Q(G) \cup D_P(G)$.

DEFINITION 3.1. *For a bipartite graph $G = (P, Q, E)$ we define the matching skeleton G' of G as the union of pairs $(S_j, T_j), j = -\infty, \ldots, +\infty$, with corresponding (fractional) matchings M_j. Note that G' contains at most $|P| + |Q| - 1$ edges.*

As before, we can show the following:

LEMMA 3.3. *1. For each $U \subseteq S_j$, one has $|T_j \cap \Gamma_{G'}(U)| \geq (1/\alpha_j)|U|$.*

2. For every $k > 0$, $\left(\left(P \setminus \bigcup_{j \geq k} S_j\right) \times \left(\bigcup_{j \geq k} T_j\right)\right) \cap E = \emptyset$, and $\left(\left(Q \setminus \bigcup_{j \leq -k} S_j\right) \times \left(\bigcup_{j \leq -k} T_j\right)\right) \cap E = \emptyset$.

Proof. Follows by construction of G'.

We note that the formulation of property (2) in Lemma 3.3 is slightly different from property (2) in Lemma 3.2. However, one can see that these formulations are equivalent when there are no (S_j, T_j) pairs for negative j, as is the case in Lemma 3.2.

4 $O(n)$ communication protocol for $CC(\frac{1}{3}, n)$

In this section, we prove that for any two bipartite graphs G_1, G_2, the maximum matching in the graph $G_1' \cup G_2$ is at least 2/3 of the maximum matching in $G_1 \cup G_2$, where G_1' is the matching skeleton of G_1. Thus, $CC(\epsilon, n) = O(n)$ for all $\epsilon \geq 1/3$; Alice sends the matching skeleton G_A' of her graph, and Bob computes a maximum matching in the graph $G_A' \cup G_B$.

Before proceeding, we establish some notation used for the next several sections. Denote by $(S_j, T_j), j = -\infty, \ldots, +\infty$ the set of pairs from the definition of G'. Recall that $S_j \subseteq P$ when $j \geq 0$ and $S_j \subseteq Q$ when $j < 0$. Also, given a maximum matching M in a bipartite graph $G = (P, Q, E)$, a *saturating cut* corresponding to M is a pair of disjoint sets $(A_1 \cup B_1, A_2 \cup B_2)$ such that $A_1 \cup A_2 = P, B_1 \cup B_2 = Q$, all vertices in $A_2 \cup B_1$ are matched by M, there are no matching edges between A_2 and B_1, and no edges at all between A_1 and B_2. The existence of a saturating cut follows from the max-flow min-cut theorem. Let ALG denote the size of the maximum matching in $G_1' \cup G_2$ and let OPT denote the size of the maximum matching in $G_1 \cup G_2$.

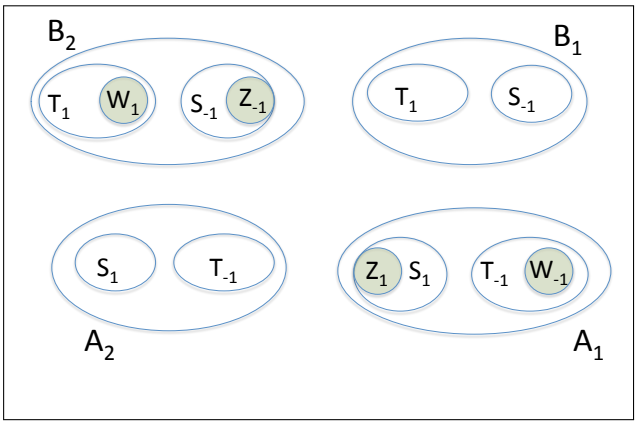

Figure 1: Distribution of (S_j, T_j) pairs across the cut

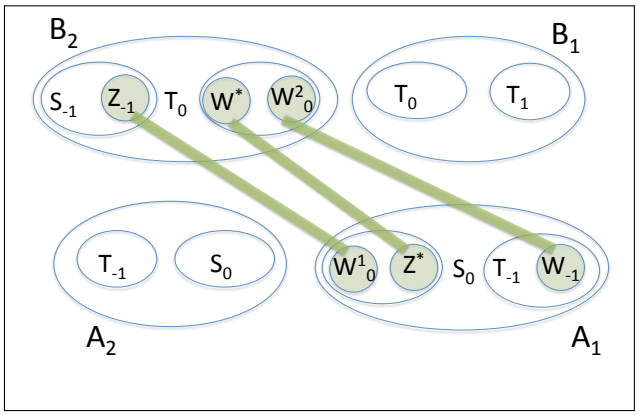

Figure 2: Matching of the (S_0, T_0) pair

Consider a maximum matching M in $(G'_1 \cup G_2)$ and a corresponding saturating cut $(A_1 \cup B_1, A_2 \cup B_2)$; note that $ALG = |B_1| + |A_2|$. Let M^* be a maximum matching in $E_1 \cap (A_1 \times B_2)$. Note that we have $OPT \leq |B_1| + |A_2| + |M^*|$.

We start by describing the intuition behind the proof. Suppose for simplicity that the matching skeleton G'_1 of G_1 consists of only one (S_j, T_j) pair for some $j \geq 0$, such that $|T_j| = (1/\alpha_j)|S_j|$. We first note that since the matching M^* is not part of the matching skeleton, it must be that edges of M^* go from S_j to T_j. We will abuse notation slightly by writing $M^* \cap X$ to denote, for $X \subseteq P \cup Q$, the subset of nodes of X that are matched by M^*. Since all edges of M^* go from S_j to T_j, we have $M^* \cap A_1 \subseteq S_j \cap A_1$ and $M^* \cap B_2 \subseteq T_j \cap B_2$. This allows us to obtain a lower bound on $|B_1|$ and $|A_2|$ in terms of $|M^*|$ if we lower bound $|B_1|$ and $|A_2|$ in terms of $|S_j \cap A_1|$ and $|T_j \cap B_2|$ respectively. First, we have that $|B_1| \geq |\Gamma_{G'_1}(S_j \cap A_1)| \geq (1/\alpha_j)|S_j \cap A_1| \geq (1/\alpha_j)|M^*|$, where we used the fact that the saturating cut is empty in $G'_1 \cup G_2$ and Lemma 3.3 . Next, we prove that $|\Gamma_{G'_1}(S_j \cap A_2) \cap B_2| \leq (1/\alpha_j)|S_j \cap A_2|$ (this is proved in Lemma 4.2 below). This, together with the fact that $M^* \cap B_2 \subseteq T_j \cap B_2 = \Gamma_{G'_1}(S_j \cap A_2) \cap B_2$, implies that $|A_2| \geq \alpha_j |M^*|$. Thus, we always have $|A_2| + |B_1| \geq (\alpha_j + 1/\alpha_j)|M^*|$, and hence the worst case happens at $\alpha_j = 1$, i.e. when the matching skeleton G'_1 of G_1 consists of only the (S_0, T_0) pair, yielding a 2/3 approximation. The proof sketch that we just gave applies when the matching skeleton only contains one pair (S_j, T_j). In the general case, we use Lemma 3.3 to control the distribution of M^* among different (S_j, T_j) pairs. More precisely, we use the fact that edges of M^* may go from $S_j \cap A_1$ to $T_i \cap B_2$ *only if* $i \leq j$. Another aspect that adds complications to the formal proof is the presence of (S_j, T_j) pairs for negative j.

We will use the notation

$$Z_j \subseteq \begin{cases} S_j \cap A_1, & j > 0 \\ S_j \cap B_2, & j < 0. \end{cases}$$

and

$$W_j \subseteq \begin{cases} T_j \cap B_2, & j > 0 \\ T_j \cap A_1, & j < 0 \end{cases}$$

for the vertices in P and Q that are matched by M^* (see Fig. 1). Further, let Z^* denote the set of vertices in $S_0 \cap A_1$ that are matched by M^* to $B_2 \cap T_0$, and let $W^* = M^*(Z^*) \subseteq B_2 \cap T_0$. Let $W_0^1 \subseteq S_0 \cap A_1$ denote the vertices in $S_0 \cap A_1$ that are matched by M^* outside of T_0. Similarly, let $W_0^2 \subseteq T_0 \cap B_2$ denote the vertices in $T_0 \cap B_2$ that are matched by M^* outside of S_0 (see Fig. 2). Let

$$B'_1 := B_1 \cap \left(\Gamma_{G'_1}(Z^*) \cup \Gamma_{G'_1}(W_0^1) \cup \bigcup_{j>0} \left(\Gamma_{G'_1}(Z_j) \cup S_{-j} \right) \right)$$

$$A'_2 := A_2 \cap \left(\Gamma_{G'_1}(W^*) \cup \Gamma_{G'_1}(W_0^2) \cup \bigcup_{j<0} \left(\Gamma_{G'_1}(Z_j) \cup S_{-j} \right) \right).$$

Then since

$$OPT \leq |B'_1| + |A'_2| + |M^*| + (|B_1 \setminus B'_1| + |A_2 \setminus A'_2|)$$
$$ALG = |B'_1| + |A'_2| + (|B_1 \setminus B'_1| + |A_2 \setminus A'_2|),$$

it is sufficient to prove that $(|B'_1| + |A'_2|) \geq (2/3)(|B'_1| + |A'_2| + |M^*|)$. Let $OPT' = |B'_1| + |A'_2| + |M^*|$ and $ALG' = |B'_1| + |A'_2|$. Define $\Delta' = (OPT' - ALG')/OPT'$.

We will now define variables to represent the sizes

of the sets used in defining B'_1, A'_2:

$$w_0^1 = |W_0^1|, w_0^2 = |W_0^2|,$$
$$z^* = |Z^*|, w^* = |W^*|, (\text{Note that } z^* = w^*)$$
$$z_j = |Z_j|, w_j = |W_j|, r_j = |\Gamma_{G'_1}(Z_j)|,$$
$$s_j = \begin{cases} |S_j \cap A_2| & j > 0 \\ |S_j \cap B_1| & j < 0 \end{cases}$$

Lemma 4.1 expresses the size of B'_1 and A'_2 in terms of the new variables defined above.

LEMMA 4.1. *$ALG' = \sum_{j\neq 0}(s_j + r_j) + (z^* + w_0^1) + (w^* + w_0^2)$, and $OPT' \leq z^* + (z^* + w_0^1) + (w^* + w_0^2) + \sum_{j\neq 0}(s_j + z_j + r_j)$.*

Proof. The main idea is that most of the sets in the definitions of B'_1 and A'_2 are disjoint, allowing us to represent sizes of unions of these sets by sums of sizes of individual sets.

For ALG', recall that $\Gamma_{G'_1}(S_j) = T_j$ and hence, the sets $\Gamma_{G'_1}(S_j)$ are all disjoint. Further, the sets S_j are all disjoint, by construction, and disjoint with all the T_j's. Thus, $|A'_1| + |B'_2| = |\Gamma_{G'_1}(W^*) \cup \Gamma_{G'_1}(W_0^2)| + |\Gamma_{G'_1}(Z^*) \cup \Gamma_{G'_1}(W_0^1)| + \sum_{j\neq 0}(s_j + r_j)$. The sets W^* and W_0^2 are disjoint. Further, they are subsets of T_0 (corresponding to $\alpha = 1$), and hence nodes in these sets have a single unique neighbor in G'_1; consequently $|\Gamma_{G'_1}(W^*) \cup \Gamma_{G'_1}(W_0^2)| = w^* + w_0^2$. Similarly, $|\Gamma_{G'_1}(Z^*) \cup \Gamma_{G'_1}(W_0^1)| = z^* + w_0^1$. This completes the proof of the lemma for ALG'.

We have $OPT' = ALG' + |M^*|$. Consider any edge $(u, v) \in M^*$. This edge is not in G'_1 and hence must go from an S_j to a $T_{j'}$ where $0 \leq j' \leq j$ or $0 \geq j' \geq j$. The number of edges in M^* that go from S_0 to T_0 is precisely z^* by definition; the number of remaining edges is precisely $\sum_{j\neq 0} z_j$.

We now derive linear constraints on the size variables, leading to a simple linear program. We have by Lemma 3.3 that for all $k > 0$

$$\begin{aligned} \left(\left(P \setminus \bigcup_{j\geq k} Z_j\right) \times \left(\bigcup_{j\geq k} W_j\right)\right) \cap E_1 = \emptyset, \\ \left(\left(Q \setminus \bigcup_{j\leq -k} Z_j\right) \times \left(\bigcup_{j\leq -k} W_j\right)\right) \cap E_1 = \emptyset. \end{aligned} \tag{4.1}$$

The existence of M^* together with (4.1)yields

$$\begin{aligned} \sum_{j=k}^{+\infty} z_j \geq \sum_{j=k}^{+\infty} w_j, \forall k > 0, \\ \sum_{j=-\infty}^{-k} z_j \geq \sum_{j=-\infty}^{-k} w_j, \forall k > 0. \end{aligned} \tag{4.2}$$

Furthermore, we have by definition of W_0^1 together with (4.1)that

$$\begin{aligned} w_0^1 \leq \sum_{j<0} z_j - \sum_{j<0} w_j \\ w_0^2 \leq \sum_{j>0} z_j - \sum_{j>0} w_j. \end{aligned} \tag{4.3}$$

Also, we have

$$\begin{aligned} \sum_{j<0} z_j = w_0^1 + \sum_{j<0} w_j \\ \sum_{j>0} z_j = w_0^2 + \sum_{j>0} w_j. \end{aligned} \tag{4.4}$$

Next, by Lemma 3.3, we have $r_j \geq (1/\alpha_j) z_j$. We also need

LEMMA 4.2. *(1) $|\Gamma_{G'_1}(S_j \cap A_2) \cap B_2| \leq (1/\alpha_j)|S_j \cap A_2|$ for all $j > 0$, and (2) $|\Gamma_{G'_1}(S_j \cap B_1) \cap A_1| \leq (1/\alpha_j)|S_j \cap B_1|$ for all $j < 0$.*

Proof. We prove (1). The proof of (2) is analogous. Suppose that $|\Gamma_{G'_1}(S_j \cap A_2) \cap B_2| > (1/\alpha_j)|S_j \cap A_2|$. Then using the assumption that $(A_1 \times B_2) \cap E' = \emptyset$, we get

$$\begin{aligned} |T_j| &= |T_j \cap B_2| + |T_j \cap B_1| \\ &\geq |\Gamma_{G'_1}(S_j \cap A_2) \cap B_2| + |\Gamma_{G'_1}(S_j \cap A_1)| \\ &> (1/\alpha_j)|S_j \cap A_2| + (1/\alpha_j)|S_j \cap A_1| > (1/\alpha_j)|S_j|, \end{aligned}$$

a contradiction to the definition of the matching skeleton.

We will now bound $\Delta' = (OPT' - ALG')/OPT'$ using a sequence of linear programs, described in figure 3. We will overload notation to use P_1^*, P_2^*, P_3^*, respectively, to refer to these linear programs as well as their optimum objective function value. By Lemma 4.2 one has for all $j \neq 0$ that $(1/\alpha_j)s_j \geq w_j$. We combine this with equations 4.2, 4.3, and 4.4 to obtain the first of our linear programs, P_1^*, in figure 3. Bounding Δ' is equivalent to bounding this LP (i.e. $\Delta' \leq P_1^*$). Note that we have implicitly rescaled the variables so that $OPT' \leq 1$.

We now symmetrize the LP P_1^* by collecting the variables for cases when j is positive, negative, and 0 to obtain LP P_2^* in figure 3. Finally, we relax LP P_2^* by combining the second and third constraints, and then establish that the remaining constraints are all tight. This gives us the LP P_3^* in figure 3. Details of the construction are embedded in the proof of the following lemma.

LEMMA 4.3. *$P_1^* \leq P_2^* \leq P_3^*$.*

$$P_1^* = \max\ z^* + \sum_{j\neq 0} z_j$$

s.t.

$$z^* + (z^* + w_0^1)$$
$$+(w^* + w_0^2) + \sum_{j\neq 0} s_j + z_j + r_j \leq 1$$
$$\forall k > 0, \sum_{j=k}^{+\infty} z_j \geq \sum_{j=k}^{+\infty} w_j,$$
$$\forall k > 0, \sum_{j=-\infty}^{-k} z_j \geq \sum_{j=-\infty}^{-k} w_j$$
$$\forall j \neq 0, (1/\alpha_j) s_j \geq w_j$$
$$\forall j \neq 0, r_j \geq (1/\alpha_j) z_j$$
$$\sum_{j<0} z_j = w_0^1 + \sum_{j<0} w_j$$
$$\sum_{j>0} z_j = w_0^2 + \sum_{j>0} w_j$$
$$z^* = w^*$$
$$s, z, w, r, z^*, w^*, w_0^1, w_0^2 \geq 0$$

$$P_2^* = \max \sum_{j=0}^{+\infty} z_j$$

s.t.

$$\sum_{j=0}^{+\infty} s_j + z_j + r_j \leq 1$$
$$\forall k \geq 0, \sum_{j=0}^{k} w_j \geq \sum_{j=0}^{k} z_j$$
$$(1/\alpha_j) s_j \geq w_j, j \geq 0$$
$$r_j \geq (1/\alpha_j) z_j, j \geq 0$$
$$x, z, w, r \geq 0$$

$$P_3^* = \max \sum_{j=0}^{\infty} z_j$$

s.t.

$$\sum_j (\alpha_j + 1 + 1/\alpha_j) z_j \leq 1$$
$$z \geq 0$$

Figure 3: The linear programs for lower bounding ALG/OPT.

From P_1^* to P_2^* We will show that the optimum of the LP P_2^* in figure 3 is an upper bound for the optimum of P_1^* in figure 3. First increase the set $\{\alpha_j\}_{j=-\infty}^{\infty}$ to ensure that $\alpha_j = \alpha_{-j}$(this can only improve the objective function). Now, we define

$$\begin{aligned}
s'_j &= s_j + s_{-j}, j > 0,\\
r'_j &= r_j + r_{-j}, j > 0,\\
z'_j &= z_j + z_{-j}, j > 0,\\
w'_j &= w_j + w_{-j}, j > 0,\\
w'_0 &= w^* + w_0^1 + w_0^2,\\
s'_0 &= w^* + w_0^1 + w_0^2,\\
z'_0 &= z^*,\\
r'_0 &= z^*.
\end{aligned} \tag{4.5}$$

We will show that if $s, r, z, w, z^*, w^*, w_0^1, w_0^2$ are feasible for P_1^*, then s', r', z', w' are feasible for P_2^* with the same objective function value.

First, the objective function is exactly the same by inspection. Constraints 3 and 4 of P_2^* for $j > 0$ are linear in the respective variables and are hence satisfied. Furthermore, one has

$$(1/\alpha_0) s'_0 = w^* + w_0^1 + w_0^2 = w'_0$$

and

$$r'_0 = z^* = z'_0.$$

Hence, constraints 3 and 4 are satisfied for all $j \geq 0$.

To verify that constraint 1 is satisfied, we calculate

$$\begin{aligned}
\sum_{j=0}^{+\infty} s'_j + z'_j + r'_j &= s'_0 + z'_0 + r'_0 + \sum_{j=1}^{+\infty} (s'_j + z'_j + r'_j)\\
&= (w^* + w_0^1 + w_0^2) + z^* + z^* + \sum_{j\neq 0} (s_j + z_j + r_j)\\
&= z^* + (z^* + w_0^1) + (z^* + w_0^2) + \sum_{j\neq 0} (s_j + z_j + r_j) \leq 1.
\end{aligned}$$

We now verify that constraint 2 of P_2^* is satisfied. First, for $k = 0$ one has

$$w'_0 = w^* + w_0^1 + w_0^2 \geq w^* = z^* = z'_0.$$

Next, note that by adding constraints 2,3 of P_1^* we get

$$\sum_{|j|\geq k} z_j \geq \sum_{|j|\geq k} w_j \tag{4.6}$$

for all $k > 0$. Adding constraints 6 and 7 of P_1^*, we get

$$\sum_{j\neq 0} z_j = w_0^1 + w_0^2 + \sum_{j\neq 0} w_j. \tag{4.7}$$

Subtracting (4.7) from (4.6), we get

$$\sum_{|j|=1}^{k} z_j \le w_0^1 + w_0^2 + \sum_{|j|=1}^{k} w_j. \tag{4.8}$$

Adding z^* to both sides and using the fact that $z'_0 = z^*$ and $w'_0 = z^* + w_0^1 + w_0^2$, we get

$$\sum_{j=0}^{k} z_j \le \sum_{j=0}^{k} w_j. \tag{4.9}$$

This completes the proof of the first half of lemma 4.3.

From P_2^* to P_3^* We now bound P_2^*. First we relax the constraints by adding constraint 3 of over j from 0 to k and adding to constraint 2:

$$\begin{aligned} \max \sum_{j=0}^{\infty} z_j \\ \text{s.t.} \\ \sum_{j=0}^{\infty} s_j + z_j + r_j \le 1 \\ \sum_{j=0}^{k} (1/\alpha_j) s_j \ge \sum_{j=0}^{k} z_j, \forall k \ge 0 \\ r_j \ge (1/\alpha_j) z_j, \forall j \ge 0 \\ x, z, w, r \ge 0 \end{aligned} \tag{4.10}$$

Note that the first constraint is necessarily tight at the optimum. Otherwise scaling all variables to make the constraint tight increases the objective function. We now show that all of the constraints in the second line of (4.10) are necessarily tight at the optimum. Indeed, let $k^* \ge 0$ be the smallest such that $\sum_{j=0}^{k^*} (1/\alpha_j) s_j > \sum_{j=0}^{k^*} z_j$. Note that one necessarily has $s_{k^*} > 0$. Let

$$\begin{aligned} s' &= s - \delta e_{k^*} + (\alpha_{k^*+1}/\alpha_{k^*}) \delta e_{k^*+1}, \\ r' &= r, z' = z, \end{aligned}$$

where e_j denotes the vector of all zeros with 1 in position j. Then

$$\sum_{j=0}^{k} (1/\alpha_j) s'_j \ge \sum_{j=0}^{k} z'_j$$

for all k and

$$\sum_{j=0}^{\infty} (s'_j + z'_j + r'_j) = 1 - \delta(1 - \alpha_{k^*+1}/\alpha_{k^*}).$$

So for sufficiently small positive $\delta > 0$ one has that

$$\begin{aligned} s'' &= s'/(1 - \delta(1 - \alpha_{k^*+1}/\alpha_{k^*})) \\ r'' &= r'/(1 - \delta(1 - \alpha_{k^*+1}/\alpha_{k^*})) \\ z'' &= z'/(1 - \delta(1 - \alpha_{k^*+1}/\alpha_{k^*})) \end{aligned}$$

form a feasible solution with a better objective function value.

Thus, one has $\sum_{j=0}^{k} (1/\alpha_j) s_j = \sum_{j=0}^{k} z_j$ for all $k \ge 0$ and hence $(1/\alpha_j) s_j = z_j$ for all j.

Additionally, one necessarily has $r_j = (1/\alpha_j) z_j$ for all j at optimum. Indeed, otherwise decreasing r_j does not violate any constraint and makes constraint 1 slack. Then rescaling variables to restore tightness of constraint 1 improves the objective function. Thus, we need to solve

$$\begin{aligned} P_3^* = \max \sum_{j=0}^{\infty} z_j \\ \text{s.t.} \\ \sum_j (\alpha_j + 1 + 1/\alpha_j) z_j \le 1 \\ z \ge 0 \end{aligned} \tag{4.11}$$

But P_3^* is easy to analyze: there exists an optimum solution that sets all z_j to zero except for a j that minimizes $(\alpha_j + 1 + 1/\alpha_j)$. For all non-negative x, $f(x) = 1 + x + 1/x$ is minimized when $x = 1$, and $f(1) = 3$. This gives $P_3^* \le 1/3$, and hence $\Delta' \le 1/3$, or $ALG' \ge (2/3) OPT'$. Thus, we have proved

THEOREM 4.1. *For any bipartite graph $G_1 = (P, Q, E_1)$ there exists a subforest G'_1 of G such that for any graph $G_2 = (P, Q, E_2)$ the maximum matching in $G'_1 \cup G_2$ is a 2/3-approximation of the maximum matching in $G_1 \cup G_2$; further, it suffices to choose G'_1 to be the matching skeleton of G_1.*

COROLLARY 4.1. $CC(\frac{1}{3}, n) = O(n)$.

Theorem 4.1 also implies that the matching skeleton gives a linear size 1/2-cover of G.

COROLLARY 4.2. *For any bipartite graph $G = (P, Q, E)$, the matching skeleton G' is a $\frac{1}{2}$-cover of G.*

Proof. We need to show that for any $A \subseteq P, B \subseteq Q, |A|, |B| > n/2$ such that there exists a perfect matching between A and B in G one has $E' \cap (A \times B) \neq \emptyset$. Let $G_2 = (P \cup P', Q \cup Q', M_P \cup M_Q)$ be a graph that consists of a perfect matching from a new set of vertices P' to $Q \setminus B$ and a matching from a new set of vertices Q' to $P \setminus A$. Then the maximum matching in $G \cup G_2$ is of size $(3/2)n$.

By the max-flow min-cut theorem, the size of the matching in $G' \cup G_2$ is no larger than $|P \setminus A| + |Q \setminus B| + |E' \cap (A \times B)|$. By Theorem 4.1 the approximation ratio is at least 2/3, and $|P \setminus A| + |Q \setminus B| < n$, so it must be that $|E' \cap (A \times B)| > 0$.

5 $O(n)$ communication protocol for $CC_v(\frac{1}{4}, n)$

In this section we prove that $CC_v(\epsilon, n) = O(n)$ for all $\epsilon < 1/4$. In particular, we show that given a bipartite graph $G_1 = (P_1, Q, E_1)$, there exists a forest $F \subseteq E_1$ such that for any $G_2 = (P_2, Q, E)$ that may share nodes on the Q side with G_1 but not on the P side, the maximum matching in $G_1' \cup G_2$ is a 3/4-approximation of the maximum matching in $G_1 \cup G_2$. The broad outline of the proof is similar to the previous section, but we can now assume a special optimal matching using the assumption that G_2 may only share nodes with G_1 on the Q side. The proof uses the simple lemma below; we state it here since it is also needed in section 6.

LEMMA 5.1. *Let $G = (P, Q, E)$ be a bipartite graph and let $S \subseteq P$ be such that $|\Gamma(U)| \geq |U|$ for all $U \subseteq S$. Then there exists a maximum matching in G that matches all vertices of S.*

The proof is quite simple: start with an arbitrary maximum matching and repeatedly find and apply even length augmenting paths originating from unmatched nodes in S and going to matched nodes in $P \setminus S$, to reduce the number of unmatched nodes in S. These paths exist by our condition on S. The details are deferred to the full version of the paper.

We now state the main theorem of this section. The proof is deferred to the full version of the paper.

THEOREM 5.1. *Let $G_1 = (P_1, Q, E_1), G_2 = (P_2, Q, E_2)$ be bipartite graphs that share the vertex set on one side. Let G_1' be the matching skeleton of G_1. Then the maximum matching in $G_1' \cup G_2$ is a 3/4-approximation of the maximum matching in $G_1 \cup G_2$.*

6 One-pass streaming with vertex arrivals

Let $G_i = (P_i, Q, E_i)$ be a sequence of bipartite graphs, where $P_i \cap P_j = \emptyset$ for $i \neq j$. For a graph G, we denote by SPARSIFY$^*(G)$ the matching skeleton of G modified as follows: for each pair $(S_j, T_j), j < 0$ keep an arbitrary matching of S_j to a subset of T_j, discarding all other edges, and collect all these matchings into the (S_0, T_0) pair. Note that we have $S_j \subseteq P$, where P is the side of the graph that arrives in the stream. We have

LEMMA 6.1. *Let $G = (P, Q, E)$ be a bipartite graph. Let $G' = SPARSIFY^*(G)$. Let $(S_j, T_j), j = 0, \ldots, +\infty$ denote the set of expanding pairs. Then $E \cap (S_i \times T_j) = \emptyset$ for all $i < j$.*

Let

$$\begin{aligned} &G_1' = \text{SPARSIFY}^*(G_1) \\ &\text{and} \\ &G_i' = \text{SPARSIFY}^*(G_{i-1}' \cup G_i), i > 1 \end{aligned} \tag{6.12}$$

We will show that for each $\tau > 0$ the maximum matching in G_τ' is at least a $1 - 1/e$ fraction of the maximum matching in $\bigcup_{i=1}^{\tau} G_i$. We will slightly abuse notation by denoting the set of expanding pairs in G_τ' by $(S_\alpha(\tau), T_\alpha(\tau))$. Recall that we have $\alpha \in (0, 1]$, and $|S_\alpha(\tau)| = \alpha |T_\alpha(\tau)|$. We need the following

DEFINITION 6.1. *For a vertex $u \in P$ define its* level *after time τ, denoted by $\alpha_u(\tau)$, as the value of α such that $u \in S_\alpha(\tau)$. Similarly, for a vertex $v \in Q$ define its* level *after time τ, denoted by $\alpha_v(\tau)$, as the value of α such that $u \in T_\alpha(\tau)$. Note that for a vertex u is at level $\alpha = \alpha_u(\tau)$ the expansion of the pair $(S_\alpha(\tau), T_\alpha(\tau))$ that it belongs to is $1/\alpha$.*

Before describing the formal proof, we give an outline of the main ideas. In our analysis, we track the structure of the matching skeleton maintained by the algorithm over time. For the purposes of our analysis, at each time τ, every vertex is characterized by two numbers: its *initial level* β when it first appeared in the stream and its *current level* α at time τ (we denote the set of such vertices at time τ by $S_{\alpha,\beta}(\tau)$). Informally, we first deduce that the matching edges that our algorithm misses may only connect a vertex in $S_{\alpha,\beta}(\tau)$ to a vertex in $T_{\beta'}(\tau)$ for $\beta' \geq \beta$, and hence we are interested in the distribution of vertices among the sets $S_{\alpha,\beta}(\tau)$. We show that vertices that initially appeared at lower levels and then migrated to higher levels are essentially the most detrimental to the approximation ratio. However, we prove that for every $\lambda \in (0, 1]$, which can be thought of as a 'barrier', the number of vertices that initially appeared at level $\beta < \lambda$ but migrated to a level $\alpha \geq \lambda$ can never be larger than $\lambda \left| \bigcup_{\gamma \in [\lambda, 1]} T_\gamma(\tau) \right|$ at any time τ. This leads to a linear program whose optimum lower bounds the approximation ratio, and yields the $(1 - 1/e)$ approximation guarantee.

LEMMA 6.2. *For all $u \in P$ and for all τ, $\alpha_u(\tau + 1) \geq \alpha_u(\tau)$. Similarly for $v \in Q$, $\alpha_v(\tau + 1) \geq \alpha_v(\tau)$.*

Proof. We prove the statement by contradiction. Let τ be the smallest such that $\exists \alpha \in (0, 1]$ such that $R := \{u \in P : u \in S_\alpha(\tau), \alpha_u(\tau + 1) < \alpha_u(\tau)\} \neq \emptyset$. Let $\alpha^* = \min_{u \in R} \alpha_u(\tau + 1)$ (we have $\alpha^* < \alpha$ by assumption). Let $R^* = R \cap S_{\alpha^*}(\tau + 1)$. Note that $R^* \subseteq S_\alpha(\tau)$. We have

$$|\Gamma_{G_\tau'}(R^*)| \geq |\Gamma_{G_{\tau+1}'}(R^*)| \geq (1/\alpha^*)|R^*| > (1/\alpha)|R^*|. \tag{6.13}$$

Since $|\Gamma_{G_\tau'}(S_\alpha(\tau))| = (1/\alpha)|S_\alpha(\tau)|$, (6.13) implies that $S_\alpha(\tau) \setminus R^* \neq \emptyset$. However, since $|\Gamma_{G_\tau'}(S_\alpha(\tau) \setminus R^*)| \geq (1/\alpha)|S_\alpha(\tau) \setminus R^*|$, one has

$$\Gamma_{G_\tau'}(S_\alpha(\tau) \setminus R^*) \cap \Gamma_{G_\tau'}(R^*) \neq \emptyset.$$

This, however, contradicts the assumption that $(S_\alpha(\tau) \setminus R^*) \cap S_{\alpha^*}(\tau+1) = \emptyset$ and the fact that $G'_{\tau+1} = \text{SPARSIFY}^*(G'_\tau, G_{\tau+1})$.

The same argument also proves the monotonicity of levels for $v \in Q$.

Let $S_{\alpha,\beta}(\tau)$ denote the set of vertices in $u \in P$ such that

1. $u \in S_\beta(\tau')$, where τ' is the time when u arrived (i.e. $u \in P_{\tau'}$), and
2. $u \in S_\alpha(\tau)$.

Note that one necessarily has $\alpha \geq \beta$ by Lemma 6.2 for all nonempty $S_{\alpha,\beta}$.

We will need the following

LEMMA 6.3. *For all τ one has for all $\lambda \in (0,1]$*

$$\left((Q \setminus \bigcup_{\alpha\in[\lambda,1]} T_\alpha(\tau)) \times \bigcup_{\beta\in[\lambda,1]} S_{\alpha,\beta}(\tau)\right) \cap \bigcup_{t=1}^{\tau} E_t = \emptyset.$$

Proof. A vertex $u \in S_{\alpha,\beta}(\tau)$ with $\beta \geq \lambda$ that arrived at time τ_u could only have edges to $v \in T_{\lambda'}(\tau_u)$ for $\lambda' \geq \lambda$. By Lemma 6.2, such vertices v can only belong to $T_{\lambda''}(\tau)$ for some $\lambda'' \geq \lambda' \geq \beta \geq \lambda$, and the conclusion follows with the help of Lemma 6.1.

Let $t_\alpha(\tau) = |T_\alpha(\tau)|, s_{\alpha,\beta}(\tau) = |S_{\alpha,\beta}(\tau)|$. The quantities $t_\alpha(\tau), s_{\alpha,\beta}(\tau)$ are defined for $\alpha, \beta \in D = \{\Delta k : 0 < k \leq 1/\Delta\}$, where $1/\Delta$ is a sufficiently large integer (note that all relevant values of α, β are rational with denominators bounded by n). In what follows all summations over levels are assumed to be over the set D. Then

LEMMA 6.4. *For all τ and for all $\alpha \in (0,1]$, the quantities $t_\alpha(\tau), s_{\alpha,\beta}(\tau)$ satisfy*

$$\sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} s_{\beta,\delta}(\tau) \leq (\alpha-\Delta) \sum_{\beta\in[\alpha,1]} t_\beta(\tau). \tag{6.14}$$

Proof. The proof is by induction on τ.

Base: $\tau = 0$ At $\tau = 0$ the lhs is zero, so the relation is satisfied.

Inductive step: $\tau \to \tau+1$ Fix $\alpha \in (0,1)$. For all $\gamma \in (0, \alpha-\Delta]$ let

$$R_\gamma(\tau) = S_\gamma(\tau) \cap \left(\bigcup_{\beta\in[\alpha,1]} S_\beta(\tau+1)\right).$$

We have $|\Gamma_{G'_\tau}(R_\gamma(\tau))| \geq (1/\gamma)|R_\gamma(\tau)|$ and $\Gamma_{G'_\tau}(R_\gamma(\tau)) \subseteq \bigcup_{\beta\in[\alpha,1]} T_\beta(\tau+1)$.

Also, we have by Lemma 6.2 that

$$\left(\bigcup_{\beta\in[\alpha,1]} T_\beta(\tau)\right) \cup \left(\bigcup_{\gamma\in(0,\alpha-\Delta]} \Gamma_{G'_\tau}(R_\gamma(\tau))\right) \subseteq \bigcup_{\beta\in[\alpha,1]} T_\beta(\tau+1).$$

Moreover, since $\Gamma_{G'_\tau}(R_\gamma(\tau))$ are disjoint for different γ and disjoint from $T_\beta(\tau), \beta \in [\alpha,1]$, letting $r_\gamma(\tau) = |R_\gamma(\tau)|$, we have

$$\begin{aligned}\sum_{\beta\in[\alpha,1]} t_\beta(\tau+1) &\geq \sum_{\beta\in[\alpha,1]} t_\beta(\tau) + \sum_{\gamma\in(0,\alpha-\Delta]} \frac{1}{\gamma} r_\gamma(\tau) \\ &\geq \sum_{\beta\in[\alpha,1]} t_\beta(\tau) + \frac{1}{\alpha-\Delta} \sum_{\gamma\in(0,\alpha-\Delta]} r_\gamma(\tau).\end{aligned} \tag{6.15}$$

Furthermore, by Lemma 6.2

$$\begin{aligned}&\sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} s_{\beta,\delta}(\tau+1) \\ &= \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} s_{\beta,\delta}(\tau) + \sum_{\gamma\in(0,\alpha-\Delta]} r_\gamma(\tau)\end{aligned} \tag{6.16}$$

Since by inductive hypothesis

$$\sum_{\beta\in[\alpha,1]} t_\beta(\tau) \geq \frac{1}{\alpha-\Delta} \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} s_{\beta,\delta}(\tau). \tag{6.17}$$

we have by combining (6.15), (6.16) and (6.17)

$$\begin{aligned}&\sum_{\beta\in[\alpha,1]} t_\beta(\tau+1) \\ &\geq \sum_{\beta\in[\alpha,1]} t_\beta(\tau) + \frac{1}{\alpha-\Delta} \sum_{\gamma\in(0,\alpha-\Delta]} r_\gamma(\tau) \\ &\geq \frac{1}{\alpha-\Delta} \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} s_{\beta,\delta}(\tau) \\ &+ \frac{1}{\alpha-\Delta} \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} (s_{\beta,\delta}(\tau+1) - s_{\beta,\delta}(\tau)) \\ &= \frac{1}{\alpha-\Delta} \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} s_{\beta,\delta}(\tau+1).\end{aligned}$$

In what follows we only consider sets $S_{\alpha,\beta}(\tau), T_\alpha(\tau)$ for fixed τ, and omit τ for brevity. Let $S = \bigcup_{\alpha,\beta} S_{\alpha,\beta}$. Choose a maximum matching M in G_τ that matches all of S, as guaranteed by Lemma 5.1. Let γ denote

the number of vertices in T_1 that are matched outside of S by M (note that no vertices of $T_\alpha, \alpha \in (0,1)$ are matched outside of S by lemma 6.3). For each $\alpha \in (0,1]$ let $r_\alpha \leq t_\alpha$ denote the number of vertices in T_α that are not matched by M. Then the following is immediate from lemma 6.3.

LEMMA 6.5. *For all* $\lambda \leq 1$

$$\sum_{\alpha\in[\lambda,1]} t_\alpha \geq \sum_{\alpha\in[\lambda,1],\beta\in[\lambda,1]} s_{\alpha,\beta} + \sum_{\alpha\in[\lambda,1]} r_\alpha + \gamma. \tag{6.18}$$

Proof. Follows from Lemma 6.3.

We also have

$$\sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,1]} s_{\beta,\delta} = \sum_{\beta\in[\alpha,1]} \beta t_\beta \tag{6.19}$$

for all $\alpha \in (0,1]$. By Lemma 6.4 and Lemma 6.5, we get

$$\begin{aligned} ALG &= \sum_{\alpha\in(0,1)} (t_\alpha - r_\alpha) + (t_1 - r_1 - \gamma) \\ OPT &= ALG + \gamma \\ t_1 &\geq \gamma + r_1. \end{aligned}$$

Thus, we need to minimize ALG/OPT subject to $t_1 \geq r_1 + \gamma, t_\alpha, s_{\alpha,\beta} \geq 0$ and

(6.20)

$$\begin{aligned} &\forall \alpha \in (0,1] : \sum_{\beta\in[\alpha,1]} t_\beta \geq \gamma + \sum_{\beta\in[\alpha,1]} \left(r_\beta + \sum_{\delta\in[\alpha,1]} s_{\beta,\delta} \right) \\ &\forall \alpha \in (0,1] : \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} s_{\beta,\delta} \leq (\alpha - \Delta) \sum_{\beta\in[\alpha,1]} t_\beta \\ &\forall \alpha \in (0,1] \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,1]} s_{\beta,\delta} = \sum_{\beta\in[\alpha,1]} \beta t_\beta. \end{aligned}$$

We start by simplifying (6.20). First note that we can assume without loss of generality that $r_1 = 0$. Indeed, if $r_1 > 0$, we can decrease r_1 to 0 and increase γ to keep ALG constant, without violating any constraints, only increasing OPT. Furthermore, we have wlog that $t_1 > 0$ since otherwise $ALG/OPT = 1$. Finally, note that setting $t_1 = \gamma$ only makes the ratio ALG/OPT smaller, so it is sufficient to lower bound $\sum_{\alpha\in(0,1)}(t_\alpha - r_\alpha)$ in terms of γ, and for this purpose we can set $\gamma = 1$ since this only fixes the scaling of all variables. Thus, it is sufficient to lower bound the optimum of (6.21), obtaining a lower bound of $\frac{P_1^*}{P_1^*+1}$ on the ratio ALG/OPT.

Combining constraints 2 and 3 of (6.21), we get

$$\sum_{\beta=\alpha}^{1} (1 + \alpha - \Delta) t_\beta \geq \gamma + \sum_{\beta=\alpha}^{1} \beta t_\beta.$$

(6.21)

$$\begin{aligned} P_1^* = \min & \sum_{\alpha\in(0,1)} (t_\alpha - r_\alpha) \\ & \text{s.t.} \\ \forall \alpha \in (0,1] : & \sum_{\beta\in[\alpha,1]} t_\beta \geq 1 + \sum_{\beta\in[\alpha,1]} \left(r_\beta + \sum_{\delta\in[\alpha,1]} s_{\beta,\delta} \right) \\ \forall \alpha \in (0,1] : & \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} s_{\beta,\delta} \leq (\alpha - \Delta) \sum_{\beta\in[\alpha,1]} t_\beta \\ \forall \alpha \in (0,1] & \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,1]} s_{\beta,\delta} = \sum_{\beta\in[\alpha,1]} \beta t_\beta \\ & t_\alpha, s_{\alpha,\beta} \geq 0. \end{aligned}$$

Thus, it is sufficient to lower bound the optimum of

(6.22)

$$\begin{aligned} P_2^* = \min & \sum_{\alpha\in(0,1)} (t_\alpha - r_\alpha) \\ & \text{s.t.} \\ \forall \alpha \in (0,1] : & \sum_{\beta\geq\alpha} (1 - \beta + \alpha - \Delta) t_\beta \geq 1 + \sum_{\beta\in[\alpha,1)} r_\alpha. \\ & t_\alpha \geq 0. \end{aligned}$$

We first show that one has $r_\alpha = 0$ for all $\alpha \in [0,1)$ at the optimum. Indeed, suppose that $r_{\alpha^*} > 0$ for some $\alpha^* \in (0,1)$. Then since the coefficient of t_{α^*} is $(1 - \alpha^* + \alpha - \Delta) \leq 1 - \Delta < 1$, $\beta = \alpha^* \geq \alpha$, we can decrease r_α^* by some $\delta > 0$ and also decrease t_{α^*} by $\frac{\delta}{1-\Delta} < \delta$, keeping all constraints satisfied and improving the value of the objective function.

Thus, we arrive at the final LP, whose optimum we need to lower bound:

(6.23)

$$\begin{aligned} P_3^* = \min & \sum_{\alpha\in(0,1)} t_\alpha \\ & \text{s.t.} \\ \forall \alpha \in (0,1] : & \sum_{\beta\geq\alpha} (1 - \beta + \alpha - \Delta) t_\beta \geq 1. \\ & t_\alpha \geq 0. \end{aligned}$$

We now show that all constraints are necessarily tight at the optimum. Let $\alpha^* \in [0,1]$ be the largest such that constraint 1 is not tight. Note that one necessarily

has $t_{\alpha^*} > 0$. Let $t' = t - \delta e_{\alpha^*} + \frac{\delta}{1+\Delta} e_{\alpha^*-\Delta}$.

We now verify that all constrains are satisfied. For $\alpha > \alpha^*$ all constraints are satisfied since we did not change t. For $\alpha = \alpha^*$, the constraint is satisfied since it was slack for t and δ is sufficiently small.

For $\alpha < \alpha^*$, i.e. $\alpha \leq \alpha^* - \Delta$ since we are considering only $\alpha \in D$, we have

$$\begin{aligned}\sum_{\beta \geq \alpha} (1 - \beta + \alpha - \Delta) t'_\beta &= \sum_{\beta \geq \alpha} (1 - \beta + \alpha - \Delta) t_\beta \\ + \delta \left(\frac{1 - (\alpha^* - \Delta) + \alpha - \Delta}{1 + \Delta} - (1 - \alpha^* + \alpha - \Delta) \right) &= \sum_{\beta \geq \alpha} (1 - \beta + \alpha - \Delta) t_\beta + \frac{\delta \Delta (\alpha^* - \alpha - \Delta)}{1 + \Delta} \\ &\geq \sum_{\beta \geq \alpha} (1 - \beta + \alpha - \Delta) t_\beta \geq 1.\end{aligned}$$

Thus, at the optimum we have

$$\sum_{\beta \geq \alpha} (1 + (\alpha - \beta - \Delta)) t_\beta = 1, \forall \alpha \in [0, 1]. \tag{6.24}$$

Subtracting (6.24) for $\alpha + \Delta$ from (6.24) for α, we get

$$\begin{aligned}&\sum_{\beta \geq \alpha} (1 + (\alpha - \beta - \Delta)) t_\beta \\ &- \sum_{\beta \geq \alpha + \Delta} (1 + (\alpha + \Delta - \beta - \Delta) t_\beta \\ &= t_\alpha - \Delta \sum_{\beta \geq \alpha} t_\beta = 0.\end{aligned} \tag{6.25}$$

In other words,

$$t_\alpha = \Delta \sum_{\beta \geq \alpha} t_\beta, t_1 \geq 1. \tag{6.26}$$

Let $\delta = \frac{\Delta}{1-\Delta}$. We now prove by induction that $t_{1-k\Delta} = \delta(1+\delta)^{k-1}$ for all $k > 0$.

Base: $k = 1$ $t_{1-\Delta} = \frac{\Delta}{1-\Delta} = \delta$.

Inductive step: $k \to k + 1$

$$t_{1-(k+1)\Delta} = \Delta \left(t_{1-(k+1)\Delta} + 1 + \delta \sum_{j=1}^{k} (1+\delta)^{j-1} \right)$$

Thus,

$$\begin{aligned}t_{1-(k+1)\Delta} &= \delta \left(1 + \delta \sum_{j=1}^{k} (1+\delta)^{j-1} \right) \\ &= \delta \left(1 + \delta \frac{1 - (1+\delta)^k}{1 - (1+\delta)} \right) = \delta (1+\delta)^k.\end{aligned}$$

Hence, one has

$$\begin{aligned}\sum_{\alpha \in [0,1)} t_\alpha &\geq \delta \sum_{j=1}^{1/\Delta} (1+\delta)^{j-1} = \delta \frac{1 - (1+\delta)^{1/\Delta}}{1 - (1+\delta)} \\ &= (1+\delta)^{1/\Delta} - 1 = \left(1 + \frac{\Delta}{1-\Delta} \right)^{1/\Delta} - 1 \\ &= (1-\Delta)^{-1/\Delta} - 1\end{aligned}$$

Now, the size of the matching M is bounded by

$$OPT \leq \sum_{\alpha \in [0,1)} t_\alpha + 1.$$

On the other hand,

$$ALG \geq \sum_{\alpha \in [0,1)} t_\alpha.$$

Thus, we get

$$\begin{aligned}\frac{ALG}{OPT} = \frac{P_1^*}{P_1^* + 1} &= 1 - \frac{1}{P_1^* + 1} \geq 1 - \frac{1}{P_3^* + 1} \\ &\geq 1 - (1-\Delta)^{1/\Delta} \geq 1 - 1/e\end{aligned}$$

since $(1-\Delta)^{1/\Delta} \leq 1/e$ for all $\Delta \geq 0$. We now prove

THEOREM 6.1. *There exists a deterministic $O(n)$ space 1-pass streaming algorithm for approximating the maximum matching in bipartite graphs to factor $1 - 1/e$ in the vertex arrival model.*

Proof. Run the algorithm given in (6.12), letting $|P_i| = 1$, i.e. sparsifying as soon as a new vertex comes in. The algorithm only keeps a sparsifier G'_i in memory, which takes space $O(n)$.

7 Constructions of Ruzsa-Szemerédi graphs

In this section we give two extensions of constructions of Ruzsa-Szemerédi graphs from [7]. The first construction shows that for any constant $\epsilon > 0$ there exist $(1/2 - \epsilon)$-Ruzsa-Szemerédi graphs with superlinear number of edges. We use this construction in section 8 to prove that our bound on $CC(\epsilon, n), \epsilon < 1/3$ is tight. The second construction that we present is a generalization to lop-sided graphs, which we use in section 8 to prove that our bound on $CC_v(\epsilon, n), \epsilon < 1/4$ is tight. Specifically, we show the following results:

LEMMA 7.1. *For any constant $\epsilon > 0$ there exists a family of bipartite $(1/2 - \epsilon)$-Ruzsa-Szemerédi graphs with $n^{1+\Omega(1/\log\log n)}$ edges.*

LEMMA 7.2. *For any constant $\delta > 0$ there exists a family of bipartite Ruzsa-Szemerédi graphs $G = (X, Y, E)$*

with $|X| = n$, $|Y| = 2n$ such that (1) the edge set E is a union of $n^{\Omega_\delta(1/\log\log n)}$ induced 2-matchings $M_1, \ldots, M_k$ of size at least $(1/2 - O(\delta))|X|$, and (2) for any $j \in [1:k]$ the graph G contains a matching M_j^* of size at least $(1 - O(\delta))|X|$ that avoids $Y \setminus (M_j \cap Y)$.

The proofs of these results are based on an adaptation of Theorem 16 in [7] (see also [15]), which constructs *bipartite* 1/3-Ruzsa-Szemerédi graphs with superlinear number of edges. The main idea of the construction, use of a large family of nearly orthogonal vectors derived from known families of error correcting codes, is the same. A technical step is required to go from matchings of size 1/3 to matchings of size $1/2 - \epsilon$ for any $\epsilon > 0$. Since the result does not follow directly from [7], we give a complete proof in the full version.

8 Lower bounds on communication and one-pass streaming complexity

We show here that lower bounds on the size of Ruzsa-Szemerédi graphs yield lower bounds on the (randomized) communication complexity, and hence for one-pass streaming complexity.

In the edge model, we show that $CC\left(\frac{2(1-\epsilon)}{2-\epsilon} - \delta, (2-\epsilon)n\right) = \Omega(U_I(\epsilon, n))$ for all $\epsilon, \delta > 0$. In particular, combined with the constructions of $(1/2 + \delta_0)$-Ruzsa-Szemerédi graphs for any constant $\delta_0 > 0$ (Lemma 7.1) this proves that $CC(\epsilon, n) = n^{1+\Omega(1/\log\log n)}$ for $\epsilon < 1/3$. Thus our $O(n)$ upper bound on $CC(\frac{1}{3}, n)$ in section 4 is optimal in the sense that any better approximation requires super-linear communication. As a corollary, we also get that super-linear space is necessary to achieve better than 2/3-approximation in the one-pass streaming model.

In the vertex model, using the construction of Ruzsa-Szemerédi graphs from Lemma 7.2, we show that $CC_v(\epsilon, n) = n^{1+\Omega(1/\log\log n)}$ for all $\epsilon < 1/4$. This proves optimality of our construction in section 5, and also shows that super-linear space is necessary to achieve better than 3/4-approximation in the one-pass streaming model even in the vertex arrival setting.

We note that our lower bounds for both the edge and vertex arrival case apply to randomized algorithms. The proofs of these results appear in the full version.

8.1 Edge arrivals

LEMMA 8.1. *For any $\epsilon > 0$ and $\delta > 0$, $CC\left(\frac{2(1-\epsilon)}{2-\epsilon} - \delta, (2-\epsilon)n\right) = \Omega(U_I(\epsilon, n))$.*

Proof. For any $\delta > 0$, we will construct a distribution over bipartite graphs with $(2-\epsilon)n$ vertices on each side such that each graph in the distribution contains a matching of size at least $(2-\epsilon)n - \delta n$. On the other hand, we will define a partition of the edge set E of the graph into $E = E_1 \cup E_2$ and show that any for deterministic communication protocol using message size $s = o(U_I(\epsilon, n))$, the expected size of the matching computed is bounded by $2(1-\epsilon)n + o(n)$. Using Yao's minmax principle, we get the desired performance bound for any protocol with $o(U_I(\epsilon, n))$ communication.

Let $G = (P, Q, E)$ be an ϵ-RS graph with n vertices on each side and $U_I(\epsilon, n)$ edges. By definition, E can be partitioned into k induced matchings $M_1, ..., M_k$, where $|M_i| = \epsilon n$ for $1 \le i \le k$, and $k = U_I(\epsilon, n)/(\epsilon n)$. We generate a random bipartite graph $G' = (P_1 \cup P_2, Q_1 \cup Q_2, E_1 \cup E_2)$ with $(2-\epsilon)n$ vertices on each side, as follows:

1. We set $P_1 = P$ and $Q_1 = Q$. Also, let P_2 and Q_2 be a set of $(1-\epsilon)n$ vertices each that are disjoint from P and Q.
2. For each M_i, $i = 1, ..., k$, let M_i' be a uniformly at random chosen subset of M_i of size $(1-\delta)n$. We set $E_1 = \cup_{i=1}^k M_i'$.
3. Choose a uniformly random $r \in [1:k]$. Let M_1^* be an arbitrary perfect matching between P_2 and $Q \setminus Q_1(M_r)$, and let M_2^* be an arbitrary perfect matching between Q_2 and $P \setminus P_1(M_r)$. We set $E_2 = M_1^* \cup M_2^*$.

The instance G' is partitioned between Alice and Bob as follows: Alice is given all edges in $G_1(P_1, Q_1, E_1)$ (first phase), and Bob is given all edges in $G_2(P_2, Q_2, E_2)$ (second phase). Clearly, any optimal matching in G' has size at least $(2-\epsilon)n - \delta n$; consider, for instance, the matching $M_r' \cup M_1^* \cup M_2^*$.

We now show that for any deterministic communication protocol using communication at most $s = o(U_I(\epsilon, n))$, with probability at least $(1 - o(1))$, number of edges in M_r' retained by the algorithm at the end of the first phase is $o(n)$. Assuming this claim, we get that with probability at least $(1 - o(1))$, the size of the matching output by Bob is bounded by $2(1-\epsilon)n + o(n)$. Hence the expected size of the matching output by Bob is bounded by $2(1-\epsilon)n + o(n)$. We now establish the preceding claim.

We start by observing that the number of distinct first phase graphs is at least (assume $\delta < \epsilon/2$)

$$\binom{\epsilon n}{\delta n}^k = \binom{\epsilon n}{\delta n}^{\frac{U_I(\epsilon,n)}{\epsilon n}} = 2^{\gamma U_I(\epsilon,n)},$$

for some positive γ bounded away from 0. Let $\mathcal{G}$ denote the set of all possible first phase graphs, and let

$\phi : \mathcal{G} \to \{0,1\}^s$ be the mapping used by Alice to map graphs in $\mathcal{G}$ to a message of size $s = o(U_I(\epsilon, n))$. For any graph $H \in \mathcal{G}$, let $\Gamma(H) = \{H' \mid \phi(H') = \phi(H)\}$. Then note that for any graph $H \in \mathcal{G}$, Bob can output an edge e in the solution iff e occurs in every graph $H' \in \Gamma(H)$. For any subset F of $\mathcal{G}$, let G_F denote the unique graph obtained by intersection of all graphs in F (i.e. the graph G_F contains an edge e iff e is present in every graph in the family F).

CLAIM 8.1. *For any $0 < \epsilon' < \frac{\epsilon}{2}$ and any subset F of $\mathcal{G}$, let $I \subseteq \{1, 2, ..., k\}$ be the set of indices such that G_F contains at least $\epsilon' n$ edges from M_i for each $i \in I$. Then if $|F| \geq 2^{(\gamma - o(1))U_I(\epsilon,n)}$, $|I| = o(k)$.*

The details of the proof are deferred to the full version of the paper.

To conclude the proof, we note that a simple counting argument shows that for a uniformly at random chosen graph $H \in \mathcal{G}$, with probability at least $1 - o(1)$, we have $|\Gamma(H)| \geq 2^{(\gamma - o(1))U_I(\epsilon,n)}$. Conditioned on this event, it follows from claim 8.1 that for a randomly chosen index $r \in [1..k]$, with probability at least $1 - o(1)$, the graph $G_{\Gamma(H)}$ contains at most $\epsilon' n$ edges from M_r.

In particular, we get

COROLLARY 8.1. *For any $\delta > 0$, $CC(2/3 + \delta, n) = n^{1+\Omega_\delta(1/\log\log n)}$.*

Proof. Follows by putting together Lemma 7.1 and Lemma 8.1.

Lower bounds on communication complexity translate directly into bounds on one-pass streaming complexity:

COROLLARY 8.2. *For any constant $\delta > 0$ any (possibly randomized) one-pass streaming algorithm that achieves approximation factor $\frac{2(1-\epsilon)}{2-\epsilon} + \delta$ must use $\Omega(U_I(\epsilon, n))$ space. In particular, any one-pass streaming algorithm that achieves approximation factor $2/3 + \delta$ must use $n^{1+\Omega_\delta(1/\log\log n)}$ space.*

Proof. Follows by Lemma 7.1 and Lemma 8.1.

8.2 Vertex arrivals We now prove a lower bound on the communication complexity in the vertex arrival model using the construction of lop-sided Ruzsa-Szemerédi graphs from Lemma 7.2. The bound implies that our upper bound from section 5 is tight. Moreover, the bound yields the first lower bound on the streaming complexity in the vertex arrival model.

LEMMA 8.2. *For any constant $\delta > 0$, $CC^1_v(3/4+\delta, n) = n^{1+\Omega_\delta(1/\log\log n)}$.*

Proof. For sufficiently small $\delta > 0$, we will construct a distribution over bipartite graphs with $(2+\delta)n$ vertices on each side such that each graph in the distribution contains a matching of size at least $(2 - O(\delta))n$. On the other hand, we will show that for any deterministic protocol using space $s = n^{1+o(1/\log\log n)}$, the expected size of the matching computed is bounded by $(3/2 + O(\delta))n + o(n)$. Using Yao's minmax principle we get the desired performance bound for any $n^{1+o(1/\log\log n)}$-space randomized protocol.

Let $G = (P, Q, E)$ be an $(1/2 - \delta)$-RS graph with $|P| = n, |Q| = 2n$ and $n^{1+\Omega(1/\log\log n)}$ edges, as guaranteed by Lemma 7.2. By definition, E can be partitioned into k induced 2-matchings $M_1, ..., M_k$, where $|M_i| \geq (1/2 - \delta')n$ for $1 \leq i \leq k$, and $k = n^{\Omega(1/\log\log n)}$ and some $\delta' = O(\delta)$. We generate a random bipartite graph $G' = (P_1 \cup P_2, Q, E_1 \cup E_2)$ with $(2 + \delta')n$ vertices on each side, as follows:

1. We set $P_1 = P$ and let P_2 be a set of $(1 + \delta')n$ vertices that are disjoint from P.
2. For each M_i, $i = 1, ..., k$, let M'_i be a uniformly at random chosen subset of M_i of size $(1/2 - 2\delta')n$. We set $E_1 = \cup_{i=1}^k M'_i$.
3. Choose a uniformly random $r \in [1 : k]$. Let M^* be an arbitrary perfect matching between P_2 and $Q \setminus Q(M_r)$. We set $E_2 = M^*$.

Let Alice hold the graph $G_A(P_1, Q_1, E_1)$ and let Bob hold the graph $G_2 = (P_2, Q, E_2)$. By Lemma 7.2, there exists a matching M^*_r that matches at least a $(1 - \delta')$ fraction of X and avoids $Q \setminus Q(M_r)$. Thus, any optimal matching in $G_A \cup G_B$ has size at least $(2 - O(\delta))n$; consider, for instance, the matching $M^*_r \cup M^*$.

However, no deterministic space protocol can output more than a $\delta'' = O(\delta')$ fraction of the edges in M'_r if it uses $n^{1+o_{\delta''}(1/\log\log n)}$ space by the same argument as in 8.1. Hence, the size of the matching output by the protocol is bounded above by $(1/2 + O(\delta))|P_1| + |P_2| = (3/2 + O(\delta))n$.

We immediately get

COROLLARY 8.3. *For any constant $\delta > 0$ any (possibly randomized) one-pass streaming algorithm that achieves approximation factor $3/4 + \delta$ must use $n^{1+\Omega_\delta(1/\log\log n)}$ space.*

9 Matching covers & Ruzsa-Szemerédi graphs

In this section we prove that the size of the smallest possible matching cover is essentially the same as the number of edges in the largest Ruzsa-Szemerédi graph with appropriate parameters.

We are now ready to state the two theorems that use induced matchings to bound the size of matching covers. The lower bound is easy, and is proved first. The upper bound is more intricate, and is presented in section 9.1.

THEOREM 9.1. **[Lower bound]** *For any* $\delta > 0$, $L_C(\epsilon, n) \geq U_I\left((1+\delta)\epsilon, n\right) \cdot \left(\frac{\delta}{1+\delta}\right)$.

Proof. Let $c = 1 + \delta$. By definition, there exists an undirected bipartite graph $G = (P, Q, E)$ with $|E| = U_I(\epsilon c, n)$, $|P| = |Q| = n$, and an induced partition $\mathcal{F}$ of G such that every set in the partition is of size at least ϵcn. Consider the smallest ϵ-matching-cover H of G, and any set $F \in \mathcal{F}$. Recall that by the definition of an induced matching, the edges in F are the only edges between $P(F)$ and $Q(F)$. Since F is a matching between $P(F)$ and $Q(F)$, and the size of F is at least ϵcn, the intersection of H and F must be of size at least $|F| - \epsilon n$, which is at least $|F| \cdot \left(\frac{c-1}{c}\right)$. Summing over all sets F in the partition $\mathcal{F}$, we get that $|H| \geq |E| \cdot \left(\frac{c-1}{c}\right)$, which proves the theorem.

In particular, choosing $\delta = 1$, we get $L_C(\epsilon, n) \geq U_I(2\epsilon, n)/2$. The upper bound is more complicated; we first state a simplified version (Theorem 9.2), and then the full version (Theorem 9.3). The simple version is a corollary of the full version; the full version is proved in section 9.1.

THEOREM 9.2. **[Simplified upper bound]** *Assume* $0 < \epsilon < 2/3, 0 < \delta < 1$, *and* $\epsilon n \geq 3$. *Then,* $L_C(n, \epsilon) \leq U_I((1-\delta)\epsilon, n) \cdot O\left(\frac{\log(1/\epsilon)}{\delta(1-\delta)}\right)$.

THEOREM 9.3. **[Upper bound]** *Assume* $\epsilon n \geq 3$, *and* $0 < \delta < 1$. *Then,*

$$L_C(n, \epsilon) \leq U_I((1-\delta)\epsilon, n) \cdot \left(\frac{8\epsilon n}{\epsilon n - 1}\right) \cdot \cdot \left(1 + \log(1/\epsilon) + \frac{\log(\epsilon n)}{8\epsilon n}\right) \cdot \left(\frac{1}{\delta(1-\delta)}\right).$$

We state the full expression in the above theorem as opposed to using asymptotic notation since the constants are simple, and it is conceivable that one may choose to apply it in regimes where ϵ is arbitrarily close to 1. Choosing $\delta = 1/2$ in Theorem 9.2, we get the interesting special case, $L_C(n, \epsilon) = O(U_I(\epsilon/2, n) \log(1/\epsilon))$.

9.1 Proof of the Upper Bound We will now prove Theorem 9.3. Assume we are given an arbitrary undirected bipartite graph $G = (P, Q, E)$ with $|P| = |Q| = n$. Assume that ϵn is an integer. Also assume that ϵn is at least 3 (of course the most interesting case is when $\epsilon > 0$ is some constant). Before proceeding, we need another definition:

DEFINITION 9.1. *A pair* (A, B), *where* $A \subseteq P$ *and* $B \subseteq Q$, *is said to be "critical" if* $|A| = |B| = M_E(A, B) = \epsilon n$, *i.e.* A, B *are both of size* ϵn *and there is a perfect matching between them. Let* $\mathcal{C}$ *denote the set of all critical pairs in* G.

We will now consider a primal-dual pair of Linear Programs. By strong duality, the optimum objective value for both LPs is the same; denote this value as Z^*. We label the constraints in the primal with the corresponding variable in the dual, and vice versa, for clarity.

$$\begin{array}{lll} \text{PRIMAL:} & Z^* = \min \sum_{e \in E} x_e & \\ \text{s.t.:} & & \\ \forall (A,B) \in \mathcal{C}: & \sum_{e \in E \cap (A \times B)} x_e \geq 1 & [\lambda_{A,B}] \\ & x \geq 0 & \end{array}$$

$$\begin{array}{lll} \text{DUAL:} & Z^* = \max \sum_{(A,B) \in \mathcal{C}} \lambda_{A,B} & \\ \text{s.t.:} & & \\ \forall (e) \in E: & \sum_{\substack{(A,B) \in \mathcal{C}: \\ e \in E \cap (A \times B)}} \lambda_{(A,B)} \leq 1 & [x_e] \\ & \lambda \geq 0 & \end{array}$$

We will relate the size of an ϵ-matching-cover of G to the primal and the size of an ϵ-induced partition of G to the dual. In particular, in the next two subsections, we will prove the following two lemmas:

LEMMA 9.1. *The graph* G *has an* ϵ*-matching-cover of size at most*

$$\left(\frac{\epsilon n}{\epsilon n - 1}\right) \cdot (2\epsilon n(1 + \log(1/\epsilon)) + \log(\epsilon n)) \cdot Z^*.$$

LEMMA 9.2. *There exists a graph* $G' = (P, Q, E')$ *with* $E' \subseteq E$ *such that* $|E'| \geq Z^* \delta(1-\delta)\epsilon n/4$ *edges, and* G' *has a* $(1-\delta)\epsilon$*-induced partition. Hence,* $U_I(n, (1-\delta)\epsilon) \geq Z^* \delta(1-\delta)\epsilon n/4$.

Theorem 9.3 is immediate from these two lemmas.

9.1.1 Proof of Lemma 9.1 A set of edges $F \subseteq E$ is said to satisfy a pair (A, B) if $|F \cap (A \times B)| > 0$. We will further break down the proof of Lemma 9.1 in two parts.

LEMMA 9.3. *If F satisfies all critical pairs, then F is an ϵ-matching-cover.*

Proof. The proof is by contradiction. Suppose F satisfies all critical pairs, but there exists a pair (A, B) such that $A \subseteq P$, $B \subseteq Q$, and $M_F(A, B) < M_E(A, B) - \epsilon n$. Consider an arbitrary maximum matching in the graph $(A, B, E \cap (A \times B))$, say H. Discard all vertices from A and B that are not incident on an edge in H, to obtain $A' \subseteq A$, $B' \subseteq B$. It is still true that $M_F(A', B') < M_E(A', B') - \epsilon n$, but now we also know that $M_E(A', B') = |H| = |A'| = |B'|$. Consider the graph $G' = (A', B', F)$. By Hall's theorem, there exists a set $A'' \subseteq A'$ and another set $B'' \subseteq B'$ such that (a) $|A''| > |B''| + \epsilon n$, and (b) $|F \cap (A'' \times (B' \setminus B''))| = 0$. Since H is perfect matching in the graph (A', B', E), there must exist at least ϵn edges of H that go from A'' to $B' \setminus B''$; let H' denote an arbitrary set of ϵn edges of H that go from A'' to $B' \setminus B''$. Let C denote the endpoints of these edges in P and D denote the endpoints of these edges in Q. Then, $|C| = |D| = \epsilon n$ and there is a perfect matching between C and D in E, i.e., the pair (C, D) is critical. But there is no edge between C and D in F (by construction), and hence F does not satisfy all critical pairs, which contradicts our assumption.

LEMMA 9.4. *There exists a set F of size at most*

$$\left(\frac{\epsilon n}{\epsilon n - 1}\right) \cdot (2\epsilon n(1 + \log(1/\epsilon)) + \log(\epsilon n)) \cdot Z^*$$

that satisfies all critical pairs.

Proof. First note that the number of critical pairs is at most $\binom{n}{\epsilon n}^2 < \left(\frac{en}{\epsilon n}\right)^{2\epsilon n} = e^{2\epsilon n(1+\log(1/\epsilon))}$.

We will now define a simple randomized rounding procedure for the solution x of the primal LP. For convenience, let γ denote the quantity $(2\epsilon n(1+\log(1/\epsilon)) + \log(\epsilon n))$. For each edge e, let $\tilde{x}_e$ denote a Bernoulli random variable which takes the value 1 with probability $p_e = \min\{1, \gamma x_e\}$, and let all $\tilde{x}_e$'s be independent. Let F denote the set of edges e for which $\tilde{x}_e = 1$.

We will now define two bad events: Let ξ_1 denote the event that $|F| > \gamma Z^* \left(\frac{\epsilon n}{\epsilon n - 1}\right)$. Let ξ_2 denote the event that F does not satisfy all critical sets.

By construction, $\mathbf{E}[|F|] = \mathbf{E}[\sum_e \tilde{x}_e] \leq \gamma \sum_e x_e = \gamma Z^*$. Hence, by Markov's inequality, $\mathbf{Pr}[\xi_1] < \frac{\epsilon n - 1}{\epsilon n} = 1 - 1/(\epsilon n)$.

Fix an arbitrary critical set (A, B). If there exists an edge $e \in E \cap (A \times B)$ such that $p_e = 1$ then (A, B) is deterministically satisfied by F. Else, it must be that $p_e = \gamma x_e$ for every edge $e \in E \cap (A \times B)$, and the probability that F does not satisfy (A, B) is at most

$$\begin{aligned} &\prod_{e \in E \cap (A \times B)} (1 - \gamma x_e) \\ &\leq e^{-\gamma \sum_{e \in E \cap (A \times B)} x_e} \\ &\leq e^{-\gamma}, \end{aligned}$$

where the third line follows from the second from feasibility of the fractional solution. Using the union bound over all critical pairs, we get $\mathbf{Pr}[\xi_2] < e^{-\log(\epsilon n)} = 1/(\epsilon n)$. Using the union bound over the two bad events, we get $\mathbf{Pr}[\xi_1 \cup \xi_2] < 1$. Hence, (using the probabilistic method), there must exist a set of edges F that satisfies all critical pairs and has size at most $\left(\frac{\epsilon n}{\epsilon n - 1}\right) \cdot (2\epsilon n(1 + \log(1/\epsilon)) + \log(\epsilon n)) \cdot Z^*$.

This concludes the proof of Lemma 9.1.

9.1.2 Proof of Lemma 9.2 This proof is also via randomized rounding, this time applied to the optimum solution of the dual LP. For every relevant pair (A, B), choose $\tilde{\lambda}_{A,B}$ to be one with probability $\delta\lambda_{A,B}/2$ and 0 otherwise; further choose the values of different $\tilde{\lambda}_{A,B}$'s independently. If $\tilde{\lambda}_{A,B} = 1$ then we say that the pair (A, B) has been selected. Initialize H to be E; we will remove edges from H till the graph (P, Q, H) has an ϵ-induced partition.

Step 1: Getting an induced partition. First, fix an arbitrary perfect matching (in E) between each selected pair, and (a) remove all edges from H that do not belong to any of these perfect matchings. Then, (b) remove all edges that belong to more than one of the graphs induced by the selected pairs. Let the new set of edges be called H_1.

Step 2: Pruning small induced sets. At this point, the collection of sets of edges $\{(A \times B) \cap H_1 : \tilde{\lambda}_{A,B} = 1\}$ forms an induced partition of the graph (P, Q, H_1). The only problem is that some of the sets in this partition may be too small. We will count a selected pair (A, B) as "good" if it induces at least $(1-\delta)\epsilon n$ edges in H_1, and "bad" otherwise. Remove all edges from H_1 that are induced by a bad selected pair to obtain the set H_2. The set (P, Q, H_2) now has a $((1-\delta)\epsilon)$-induced partition. Let k denote the number of good selected pairs; then $|H_2|$ (and hence $U_I(n, (1-\delta)\epsilon)$) is at least $k(1-\delta)\epsilon n$.

We will now show that $\mathbf{Pr}[k > \delta Z^*/4] > 0$. Consider a relevant pair (A, B) with $\lambda_{A,B} > 0$. Now, $\mathbf{Pr}[\tilde{\lambda}_{A,B} = 1] = \delta\lambda_{A,B}/2$. Consider the perfect matching F chosen between this pair (arbitrarily) in step 1 and consider any edge e in this matching. This edge will not be pruned away in step 1(a). By the

feasibility constraint in the dual,

$$\sum_{(A',B')\in\mathcal{C}:(A,B)\neq(A',B'),e\in E\cap(A'\times B')} \lambda_{A',B'} < 1.$$

Hence, the probability that this edge will belong to a selected pair other than (A, B) is less than $\delta/2$. Thus, the expected number of edges in $H_1 \cap (A \times B)$ is more than $(1 - \delta/2)\epsilon n$. The maximum number of edges in $H_1 \cap (A \times B)$ is ϵn. Applying Markov's inequality to the random variable $\epsilon n - |H_1 \cap (A \times B)|$, we get:

$$\mathbf{Pr}[|H_1 \cap (A \times B)| \geq (1 - \delta)\epsilon n \mid \tilde{\lambda}_{A,B} = 1] > 1/2.$$

Multiplying with the probability that $\tilde{\lambda}_{A,B} = 1$, we obtain:

$$\mathbf{Pr}[\text{A relevant pair } (A, B) \text{ is both selected and good}] > \delta\lambda_{A,B}/4.$$

Summing over all relevant pairs (A, B), we get $\mathbf{E}[k] > \delta Z^*/4$, and hence (using the probabilistic method again), there must exist a set of choices for $\tilde{\lambda}_{A,B}$ which make $k > \delta Z^*/4$. For this choice, we know that H_2 (and hence $U_I(n, (1 - \delta)\epsilon)$) is at least $Z^*\delta(1 - \delta)\epsilon n/4$.

This concludes the proof of Lemma 9.2.

Finally, we note that an upper bound on the size of ϵ-covers directly yields an upper bound on the communication complexity of achieving an *additive* ϵn error approximation to bipartite matching, denoted by $CC_+(\epsilon, n)$.

LEMMA 9.5. $CC_+(\epsilon, n) \leq L_C(\epsilon, n)$.

Proof. Let $G_1 = (P_1, Q_1, E_1)$ denote the bipartite graph with $|P| = |Q| = n$ that Alice holds and let $G_2 = (P_2, Q_2, E_2)$ be the graph that Bob holds. Let G_1' be a ϵ-matching cover of G_1. Consider an empty cut $(A_1 \cup B_1, A_2 \cup B_2)$ corresponding to a maximum matching M' in $(G_1' \cup G_2)$, i.e. such that $|M'| = |B_1| + |A_2|$. Let M^* denote a maximum matching in $(A_1 \times B_2) \cap E_1$. Since G_1' is an ϵ-matching cover, we have that $|M^*| < \epsilon n$.

Thus, since the maximum matching M in $G_1 \cup G_2$ is bounded by $|B_1| + |A_2| + |M^*|$ we have

$$|M| - |M'| \leq (|B_1| + |A_2| + |M^*|) - (|B_1| + |A_2|) \leq \epsilon n.$$

References

[1] K. Ahn and S. Guha. Laminar families and metric embeddings: Non-bipartite maximum matching problem in the semi-streaming model. *CoRR*, abs/1104.4058, 2011.

[2] K. Ahn and S. Guha. Linear programming in the semi-streaming model with application to the maximum matching problem. *ICALP*, pages 526–538, 2011.

[3] F. A. Behrend. On sets of integers which contain no three terms in arithmetic progression. *Proc. Nat. Acad. Sci.*, 32:331–332, 1946.

[4] András A. Benczúr and David R. Karger. Approximating s-t minimum cuts in $\tilde{O}(n^2)$ time. *Proceedings of the 28th annual ACM symposium on Theory of computing*, pages 47–55, 1996.

[5] Sebastian Eggert, Lasse Kliemann, and Anand Srivastav. Bipartite graph matchings in the semi-streaming model. *ESA 2009*, pages 492–503, 2009.

[6] Joan Feigenbaum, Sampath Kannan, Andrew McGregor, Siddharth Suri, and Jian Zhang. On graph problems in a semi-streaming model. *Theor. Comput. Sci.*, 348:207–216, 2005.

[7] E. Fischer, E. Lehman, I. Newman, S. Raskhodnikova, R. Rubinfeld, and A. Samorodnitsky. Monotonicity testing over general poset domains. *STOC*, 2002.

[8] A. Goel, A. Meyerson, and S. Plotkin. Approximate majorization and fair online load balancing. *ACM Transactions on Algorithms*, 1(2):338–349, Oct 2005.

[9] T. W. Gowers. Some unsolved problems in additive/combinatorial number theory. *http://www.dpmms.cam.ac.uk/ wtg10/addnoth.survey.dvi.*

[10] J. Hastad and A. Wigderson. Simple analysis of graph tests for linearity and pcp. *Random Structures and Algorithms*, 22, 2003.

[11] D. Karger. Random sampling in cut, flow, and network design problems. *Mathematics of Operations Research (Preliminary version appeared in the Proceedings of the 26th annual ACM symposium on Theory of computing)*, 24(2):383–413, 1999.

[12] R. Karp, U. Vazirani, and V. Vazirani. An optimal algorithm for online bipartite matching. *STOC*, 1990.

[13] J. Kleinberg, Y. Rabani, and E. Tardos. Fairness in routing and load balancing. *J. Comput. Syst. Sci.*, 63(1):2–20, 2001.

[14] A. McGregor. Finding graph matchings in data streams. *APPROX-RANDOM*, pages 170–181, 2005.

[15] S. Raskhodnikova. Property testing: Theory and applications. *Ph.D. thesis*, 2003.

[16] A. Schrijver. *Combinatorial Optimization.* Springer Verlag, 2003.

[17] T. Tao and V. Vu. *Additive Combinatorics.* Cambridge University Press, 2009.