On the communication and streaming complexity of maximum bipartite matching

Ashish Goel* Michael Kapralov† Sanjeev Khanna‡

February 23, 2012

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 restriction 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 [13]. We present here the first *deterministic* one-pass streaming $(1-\frac{1}{e})$-approximation algorithm using $O(n)$ space for this setting.

*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-0635084 and IIS-0904314.

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-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 [12, 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, 11, 19]. 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, 19]. Constructions of a collection of $n^{c/\log\log n}$ induced matchings of size $n/3 - o(n)$ were given in [7, 17]. We use the ideas of [7, 17] 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 [19].

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 skelton* of the

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

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 [13]. We give a *deterministic* one-pass streaming algorithm that matches the $(1 - \frac{1}{e})$-approximation guarantee of KVV using only $O(n)$ 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 [16] 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, [1] 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 [2]. 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, 17] 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 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 *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 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 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 [18].

3.1 P is hypermatchable in G

We note that since P is *hypermatchable*, by Hall's theorem [18], 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 5 *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 $A_1 \cup A_2 \in \mathcal{R}(\alpha)$.*

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 5.
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 6 *1. For each $U \subseteq S_j$ one has $|\Gamma_{G_j}(U)| \geq (1/\alpha_j)|U|$.*

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

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 6 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.

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 is hypermatchable to $D_P(G)$, and there is a perfect matching between $C_P(G)$ and $C_Q(G)$.

Using the above partition, we can now define a matching skeleton of G using the above partition. 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 7 *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 8 *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 $\left(\left(Q \setminus \bigcup_{j \leq -k} S_j\right) \times \left(\bigcup_{j \leq -k} T_j\right)\right) \cap E = \emptyset$.

We note that the formulation of property (2) in Lemma 8 is slightly different from property (2) in Lemma 6. 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 6.

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 $G_1 \cup G_2$.

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 8 . 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 10 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 8 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

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

for the vertices in P and Q that are matched by M^* (see Fig. 2(a) in the appendix). 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(b) in the appendix). 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 9 expresses the size of B_1' and A_2' in terms of the new variables defined above.

Lemma 9 *$ALG' = \sum_{j \neq 0}(s_j + r_j) + (z^* + w_0^1) + (w^* + w_0^2)$, and $OPT' \leq z^* + (z^* + w_0^1) + (w^* + w_0^2) + \sum_{j \neq 0}(s_j + z_j + r_j)$.*

The proof is deferred to the full version. 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 8 that for all $k > 0$

$$\left(\left(P \setminus \bigcup_{j \geq k} Z_j\right) \times \left(\bigcup_{j \geq k} W_j\right)\right) \cap E_1 = \emptyset, \quad \text{and} \quad \left(\left(Q \setminus \bigcup_{j \leq -k} Z_j\right) \times \left(\bigcup_{j \leq -k} W_j\right)\right) \cap E_1 = \emptyset. \quad (1)$$

The existence of M^* together with (1)yields

$$\sum_{j=k}^{+\infty} z_j \geq \sum_{j=k}^{+\infty} w_j, \forall k > 0, \qquad \text{and} \qquad \sum_{j=-\infty}^{-k} z_j \geq \sum_{j=-\infty}^{-k} w_j, \forall k > 0. \quad (2)$$

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

$$w_0^1 \leq \sum_{j<0} z_j - \sum_{j<0} w_j \qquad \text{and} \qquad w_0^2 \leq \sum_{j>0} z_j - \sum_{j>0} w_j. \quad (3)$$

Also, we have

$$\sum_{j<0} z_j = w_0^1 + \sum_{j<0} w_j \qquad \text{and} \qquad \sum_{j>0} z_j = w_0^2 + \sum_{j>0} w_j. \quad (4)$$

Next, by Lemma 8, we have $r_j \geq (1/\alpha_j) z_j$. We also need

Lemma 10 *(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 figures 1(a)-1(c). 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 10 one has for all $j \neq 0$ that $(1/\alpha_j)s_j \geq w_j$. We combine this with equations 2, 3, and 4 to obtain the first of our linear programs, P_1^*, in figure 1(a). Bounding Δ' is equivalent to bounding this LP (i.e. $\Delta' \leq P_1^*$). Note that we have implicitly rescaled the variables so that $OPT' \leq 1$.

$$P_1^* = \text{maximize } z^* + \sum_{j\neq 0} z_j \quad \text{s.t.}$$

$$z^* + (z^* + w_0^1) + (w^* + w_0^2) + \sum_{j\neq 0} s_j + z_j + r_j \leq 1$$

$$\forall k > 0, \sum_{j=k}^{+\infty} z_j \geq \sum_{j=k}^{+\infty} w_j,$$

$$\forall k > 0, \sum_{j=-\infty}^{-k} z_j \geq \sum_{j=-\infty}^{-k} w_j$$

$$\forall j \neq 0, (1/\alpha_j)s_j \geq w_j$$

$$\forall j \neq 0, r_j \geq (1/\alpha_j)z_j$$

$$\sum_{j<0} z_j = w_0^1 + \sum_{j<0} w_j$$

$$\sum_{j>0} z_j = w_0^2 + \sum_{j>0} w_j$$

$$z^* = w^*$$

$$s, z, w, r, z^*, w^*, w_0^1, w_0^2 \geq 0$$

$$P_2^* = \text{maximize } \sum_{j=0}^{+\infty} z_j \quad \text{s.t.}$$

$$\sum_{j=0}^{+\infty} s_j + z_j + r_j \leq 1$$

$$\forall k \geq 0, \sum_{j=0}^{k} w_j \geq \sum_{j=0}^{k} z_j$$

$$(1/\alpha_j)s_j \geq w_j, j \geq 0$$

$$r_j \geq (1/\alpha_j)z_j, j \geq 0$$

$$x, z, w, r \geq 0$$

$$P_3^* = \text{maximize } \sum_{j=0}^{\infty} z_j \quad \text{s.t.}$$

$$\sum_j (\alpha_j + 1 + 1/\alpha_j)z_j \leq 1$$

$$z \geq 0$$

Figure 1: The linear programs for lower bounding ALG/OPT.

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 1(b). 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 1(c). Again, details are in full version where we also prove: $P_1^* \leq P_2^* \leq P_3^*$. 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 $P_3^* \leq 1/3$, and hence $\Delta' \leq 1/3$, or $ALG' \geq (2/3)OPT'$. Thus, we have proved

Theorem 11 *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 12 $CC(\frac{1}{3}, n) = O(n)$.

Theorem 11 also implies that the matching skelton gives a linear size $1/2$-cover of G; the proof of the following corollary is in the full version.

Corollary 13 *For any bipartite graph $G = (P, Q, E)$, the matching skeleton G' is a $\frac{1}{2}$-cover of G.*

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 14 *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.*

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

Theorem 15 *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. Let

$$G_1' = \text{SPARSIFY}^*(G_1), \text{ and } G_i' = \text{SPARSIFY}^*(G_{i-1}' \cup G_i). \tag{5}$$

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 16 *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 17 *For all $u \in P$ and for all τ, $\alpha_u(\tau+1) \geq \alpha_u(\tau)$. Similarly for $v \in Q$, $\alpha_v(\tau+1) \geq \alpha_v(\tau)$.*

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 17 for all nonempty $S_{\alpha,\beta}$. Combining this with properties of the matching skeleton construction we obtain

$$\forall\tau, \forall\lambda \in (0,1] : \left((Q \setminus \bigcup_{\alpha\in[\lambda,1]} T_\alpha(\tau)) \times \bigcup_{\beta\in[\lambda,1]} S_{\alpha,\beta}(\tau) \right) \cap \bigcup_{t=1}^{\tau} E_t = \emptyset. \tag{6}$$

Details are in the attached full version, where this statement is stated and proved as lemma 47. 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 18 *For all τ and for all $\alpha \in (0,1]$, we have $\sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} s_{\beta,\delta}(\tau) \leq (\alpha-\Delta) \sum_{\beta\in[\alpha,1]} t_\beta(\tau)$.*

The proof is by induction on τ, and an easy consequence of lemma 17; details are in the attached full version. 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 14. 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 (6)). 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 (6).

Lemma 19 *For all $\lambda \leq 1$, $\sum_{\alpha\in[\lambda,1]} t_\alpha \geq \sum_{\alpha\in[\lambda,1],\beta\in[\lambda,1]} s_{\alpha,\beta} + \sum_{\alpha\in[\lambda,1]} r_\alpha + \gamma$.*

We also have $\sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,1]} s_{\beta,\delta} = \sum_{\beta\in[\alpha,1]} \beta t_\beta$ for all $\alpha \in (0,1]$. By Lemma 18 and Lemma 19, we get

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

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

$$\begin{aligned}
&\forall\alpha \in (0,1] : \sum_{\beta\in[\alpha,1]} t_\beta \geq \gamma + \sum_{\beta\in[\alpha,1]} \sum_{\delta\in[\alpha,1]} s_{\beta,\delta} + \sum_{\beta\in[\alpha,1]} r_\beta. \\
&\forall\alpha \in (0,1] : \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} s_{\beta,\delta} \leq (\alpha-\Delta) \sum_{\beta\in[\alpha,1]} t_\beta \\
&\forall\alpha \in (0,1] \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,1]} s_{\beta,\delta} = \sum_{\beta\in[\alpha,1]} \beta t_\beta.
\end{aligned} \tag{7}$$

This is almost a linear program. After a series of simplifications of the type presented in section 4, we show that $ALG/OPT \geq P^*/(1+P^*)$, where P^* is the LP

$$\begin{aligned} P^* = \text{minimize} \sum_{\alpha \in (0,1)} t_\alpha \qquad \text{s.t.} \\ \forall \alpha \in (0,1] : \sum_{\beta \geq \alpha} (1 - \beta + \alpha - \Delta) t_\beta = 1. \\ t_\alpha \geq 0. \end{aligned} \tag{8}$$

It is relatively straightforward to show that the optimum value of this LP is at least $(1-\Delta)^{-1/\Delta} - 1$. Details of the simplification steps and the analysis are in the attached full version. Thus, we get

$$\frac{ALG}{OPT} \geq \frac{P^*}{P^*+1} = 1 - \frac{1}{P^*+1} \geq 1 - (1-\Delta)^{1/\Delta} \geq 1 - 1/e$$

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

Theorem 20 *There exists a deterministic $O(n)$ space 1-pass streaming algorithm for approximating the maximum matching in bipartite graphs in the vertex arrival model.*

Proof: Run the algorithm given in (5), 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 21 *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 22 *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 [17]), 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 $CC\left(\frac{2(1-\epsilon)}{2-\epsilon} - \delta, n\right) = \Omega(U_I(\epsilon, n/2))$ 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 21) 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 22, 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.

9 Matching covers versus 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.

The first theorem below shows that the size of the matching cover is at least as large as the size of a Ruzsa-Szemerédi graph with appropriate parameters. The proof of this result is straightforward.

Theorem 23 [Lower bound] *For any $\delta > 0$, $L_C(\epsilon, n) \geq U_I\left((1+\delta)\epsilon, n\right) \cdot \left(\frac{\delta}{1+\delta}\right)$.*

The theorem below gives a simplified version of the complementary upper bound result.

Theorem 24 [Simplified upper bound] *Assume $0 < \epsilon < 2/3, 0 < \delta < 1$, and $\epsilon n \geq 3$. Then, $L_C(\epsilon, n) \leq U_I((1-\delta)\epsilon, n) \cdot O\left(\frac{\log(1/\epsilon)}{\delta(1-\delta)}\right)$.*

The proof of the upper bound is more intricate. We describe briefly the main idea of the proof, deferring the complete details to the full version. We formulate a linear program to minimize the number of edges needed in an ϵ-cover of a given graph G, and show that if the optimal value of the fractional cover is Z^*, there exists an integral cover of size at most $\epsilon n Z^*$ (roughly speaking). On the other hand, we show using the dual linear program (whose optimum is also Z^* by strong duality), that for any $0 < \delta < 1$, the graph G contains a subgraph G' of size roughly $\epsilon n Z^*$ such that the edges of G' can be partitioned into induced matchings of size $(1-\delta)\epsilon n$. These two results together imply the upper bound in the theorem above.

Acknowledgements

We are grateful to Madhu Sudan for introducing us to the literature on Ruzsa-Szemerédi graphs.

References

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

[2] K. Ahn and S. Guha. Linear programming in the semi-streaming model with application to the maximum matching problem. *CoRR*, abs/1104.2315, 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] N. Goyal, L. Rademacher, and S. Vempala. Expanders via random spanning trees. *SODA*, 2009.

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

[12] 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.

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

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

[15] V. I. Levenstein. Upper bounds for codes with a fixed weight of vectors (in russian). *Problems of information transmission*, pages 3–12, 1971.

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

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

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

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

THE FULL VERSION WITH DETAILED PROOFS

A 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-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 [12, 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, 11, 19]. 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, 19]. Constructions of a collection of $n^{c/\log\log n}$ induced matchings of size $n/3 - o(n)$ were given in [7, 17]. We use the ideas of [7, 17] 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 [19].

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[2]. We refer to this particular $\frac{1}{2}$-cover as a *matching skelton* of the

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

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 [13]. We give a *deterministic* one-pass streaming algorithm that matches the $(1 - \frac{1}{e})$-approximation guarantee of KVV using only $O(n)$ 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 [16] 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, [1] 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 [2]. 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 B. In section C we give the construction of the *matching skeleton*, which we use later in section D to prove that $CC(1/3, n) = O(n)$, as well as show that the matching skeleton forms a $1/2$-cover. In section E we deduce using the matching skeleton that $CC_v(1/4, n) = O(n)$. In section F 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, 17] in section G. We use these extensions in section H 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 I 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.

B Preliminaries

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

Definition 25 *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 26 *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 27 *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 28 *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.

C 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 [18].

C.1 P is hypermatchable in G

We note that since P is *hypermatchable*, by Hall's theorem [18], 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 29 *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 29.
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 30 *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 30 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 [14, 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.

C.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 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 31 *(Edmonds-Gallai decomposition, [18]) 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*

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

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

Applying Edmonds-Gallai decomposition to bipartite graphs, we get

Corollary 32 *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*

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

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 31. 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 using the above partition. 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 33 *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 34 *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 8 is slightly different from property (2) in Lemma 6. 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 6.

D $O(n)$ 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 $G_1 \cup G_2$.

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 8 . 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 10 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 8 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

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

for the vertices in P and Q that are matched by M^* (see Fig. 2(a)). 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(b)). Let

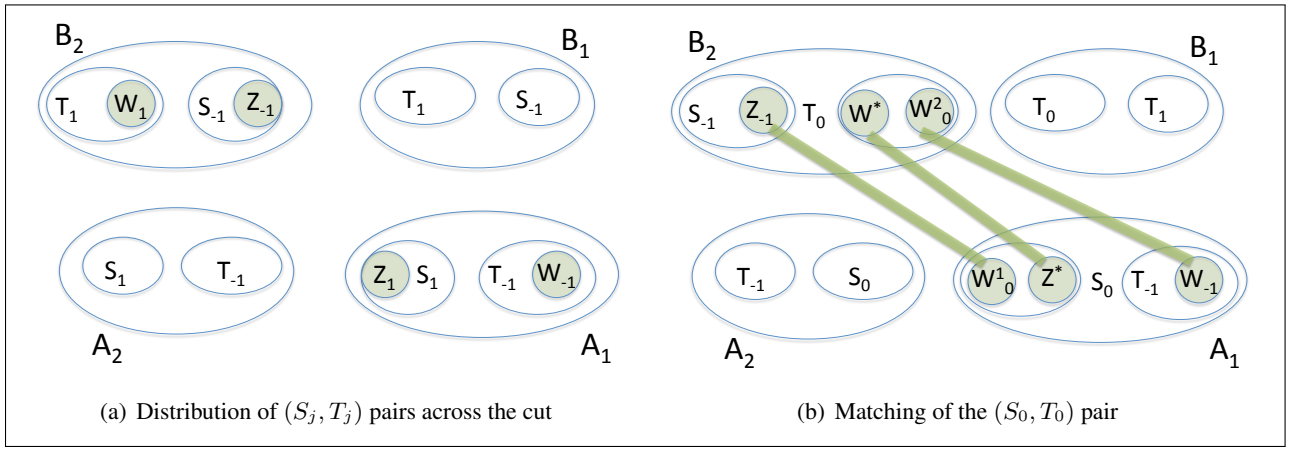
$$B_1' := B_1 \cap \left(\Gamma_{G_1'}(Z^*) \cup \Gamma_{G_1'}(W_0^1) \cup \bigcup_{j>0} \left(\Gamma_{G_1'}(Z_j) \cup S_{-j} \right) \right)$$

$$A_2' := A_2 \cap \left(\Gamma_{G_1'}(W^*) \cup \Gamma_{G_1'}(W_0^2) \cup \bigcup_{j<0} \left(\Gamma_{G_1'}(Z_j) \cup S_{-j} \right) \right).$$

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



(a) Distribution of (S_j, T_j) pairs across the cut

(b) Matching of the (S_0, T_0) pair

Figure 2: The structure of the saturating cut

.

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 35 expresses the size of B_1' and A_2' in terms of the new variables defined above.

Lemma 35 *$ALG' = \sum_{j \neq 0}(s_j + r_j) + (z^* + w_0^1) + (w^* + w_0^2)$, 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 34 that for all $k > 0$

$$\left(\left(P \setminus \bigcup_{j\geq k} Z_j\right) \times \left(\bigcup_{j\geq k} W_j\right)\right) \cap E_1 = \emptyset, \quad \text{and} \quad \left(\left(Q \setminus \bigcup_{j\leq -k} Z_j\right) \times \left(\bigcup_{j\leq -k} W_j\right)\right) \cap E_1 = \emptyset. \tag{9}$$

The existence of M^* together with (9)yields

$$\sum_{j=k}^{+\infty} z_j \geq \sum_{j=k}^{+\infty} w_j, \forall k > 0, \qquad \text{and} \qquad \sum_{j=-\infty}^{-k} z_j \geq \sum_{j=-\infty}^{-k} w_j, \forall k > 0. \tag{10}$$

Furthermore, we have by definition of W_0^1 together with (9)that

$$w_0^1 \leq \sum_{j<0} z_j - \sum_{j<0} w_j \qquad \text{and} \qquad w_0^2 \leq \sum_{j>0} z_j - \sum_{j>0} w_j. \tag{11}$$

Also, we have

$$\sum_{j<0} z_j = w_0^1 + \sum_{j<0} w_j \qquad \text{and} \qquad \sum_{j>0} z_j = w_0^2 + \sum_{j>0} w_j. \tag{12}$$

Next, by Lemma 34, we have $r_j \geq (1/\alpha_j) z_j$. We also need

Lemma 36 *(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 figures 3(a)-3(c). 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 36 one has for all $j \neq 0$ that $(1/\alpha_j)s_j \geq w_j$. We combine this with equations 10, 11, and 12 to obtain the first of our linear programs, P_1^*, in figure 3(a). Bounding Δ' is equivalent to bounding this LP (i.e. $\Delta' \leq P_1^*$). Note that we have implicitly rescaled the variables so that $OPT' \leq 1$.

$$P_1^* = \text{maximize } z^* + \sum_{j\neq 0} z_j \quad \text{s.t.}$$

$$z^* + (z^* + w_0^1) + (w^* + w_0^2) + \sum_{j\neq 0} s_j + z_j + r_j \leq 1$$

$$\forall k > 0, \sum_{j=k}^{+\infty} z_j \geq \sum_{j=k}^{+\infty} w_j,$$

$$\forall k > 0, \sum_{j=-\infty}^{-k} z_j \geq \sum_{j=-\infty}^{-k} w_j$$

$$\forall j \neq 0, (1/\alpha_j)s_j \geq w_j$$

$$\forall j \neq 0, r_j \geq (1/\alpha_j)z_j$$

$$\sum_{j<0} z_j = w_0^1 + \sum_{j<0} w_j$$

$$\sum_{j>0} z_j = w_0^2 + \sum_{j>0} w_j$$

$$z^* = w^*$$

$$s, z, w, r, z^*, w^*, w_0^1, w_0^2 \geq 0$$

$$P_2^* = \text{maximize } \sum_{j=0}^{+\infty} z_j \quad \text{s.t.}$$

$$\sum_{j=0}^{+\infty} s_j + z_j + r_j \leq 1$$

$$\forall k \geq 0, \sum_{j=0}^{k} w_j \geq \sum_{j=0}^{k} z_j$$

$$(1/\alpha_j)s_j \geq w_j, j \geq 0$$

$$r_j \geq (1/\alpha_j)z_j, j \geq 0$$

$$x, z, w, r \geq 0$$

$$P_3^* = \text{maximize } \sum_{j=0}^{\infty} z_j \quad \text{s.t.}$$

$$\sum_j (\alpha_j + 1 + 1/\alpha_j) z_j \leq 1$$

$$z \geq 0$$

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

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(b). 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(c). Details of the construction are embedded in the proof of the following lemma.

Lemma 37 $P_1^* \leq P_2^* \leq P_3^*$.

Proof:

From P_1^* to P_2^*

We will show that the optimum of the LP P_2^* in figure 3(b) is an upper bound for the optimum of P_1^* in figure 3(a). 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

$$\begin{aligned}
s'_j &= s_j + s_{-j}, j > 0 \\
r'_j &= r_j + r_{-j}, j > 0 \\
z'_j &= z_j + z_{-j}, j > 0 \\
w'_j &= w_j + w_{-j}, j > 0 \\
w'_0 &= w^* + w_0^1 + w_0^2 \\
s'_0 &= w^* + w_0^1 + w_0^2 \\
z'_0 &= z^* \\
r'_0 &= z^*.
\end{aligned} \tag{13}$$

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

$$\sum_{|j|\geq k} z_j \geq \sum_{|j|\geq k} w_j \tag{14}$$

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

$$\sum_{j\neq 0} z_j = w_0^1 + w_0^2 + \sum_{j\neq 0} w_j. \tag{15}$$

Subtracting (15) from (14), we get

$$\sum_{|j|=1}^{k} z_j \leq w_0^1 + w_0^2 + \sum_{|j|=1}^{k} w_j. \tag{16}$$

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

$$\sum_{j=0}^{k} z_j \leq \sum_{j=0}^{k} w_j. \tag{17}$$

This completes the proof of the first half of lemma 37.

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:

$$\begin{aligned} \text{maximize } & \sum_{j=0}^{\infty} z_j \\ \text{s.t.} & \\ & \sum_{j=0}^{\infty} s_j + z_j + r_j \leq 1 \\ & \sum_{j=0}^{k} (1/\alpha_j) s_j \geq \sum_{j=0}^{k} z_j, \forall k \geq 0 \\ & r_j \geq (1/\alpha_j) z_j, \forall j \geq 0 \\ & x, z, w, r \geq 0 \end{aligned} \tag{18}$$

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 (18) are necessarily tight at the optimum. Indeed, let $k^* \geq 0$ 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

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

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 $k \geq 0$ 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

$$\begin{aligned} P_3^* = \text{maximize } & \sum_{j=0}^{\infty} z_j \\ \text{s.t.} & \\ & \sum_j (\alpha_j + 1 + 1/\alpha_j) z_j \leq 1 \\ & z \geq 0 \end{aligned} \tag{19}$$

■

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 $P_3^* \leq 1/3$, and hence $\Delta' \leq 1/3$, or $ALG' \geq (2/3)OPT'$. Thus, we have proved

Theorem 38 *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 39 $CC(\frac{1}{3}, n) = O(n)$.

Theorem 38 also implies that the matching skelton gives a linear size $1/2$-cover of G.

Corollary 40 *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 38 the approximation ratio is at least $2/3$, and $|P \setminus A| + |Q \setminus B| < n$, so it must be that $|E' \cap (A \times B)| > 0$. ■

E $O(n)$ 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.

We first prove

Lemma 41 *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.*

Proof: Let M be a maximum matching in $G_1 \cup G_2$ that leaves a nonempty set $U \subseteq S$ of vertices exposed. Let U be the largest subset of S exposed by M. We will show how to obtain a different maximum matching M' that leaves one fewer nodes exposed. Orient edges of the matching M from Q to P and orient all other edges from P to Q. Denote the set of all nodes reachable from U by $\Gamma^*(U)$. Suppose that no node outside S is reachable in this directed graph. Then we have $|\Gamma^*(U) \cap Q| = |\Gamma^*(U) \cap P| - |U|$, a contradiction since

1. $\Gamma^*(U) \cap P \subseteq S$ by assumption;

2. $\Gamma^*(U) \cap Q = \Gamma(\Gamma^*(U) \cap P)$.

Thus, there exists an (even length) path in this directed graph from U to $P \setminus S$. Swapping edges in and out of M along this path decreases the number of unmatched nodes in S by one while preserving the size of the matching. Repeating the argument, we obtain a maximum matching in $G_1 \cup G_2$ that matches all of S. ■

We also need

Lemma 42 *Let $G_1 = (P_1, Q, E), G_2 = (P_2, Q, E)$ and let G_1' be the matching skeleton of G_1. Let $(A_1 \cup B_1, A_2 \cup B_2)$ be a saturating cut corresponding to a maximum matching in $G_1' \cup G_2$. Then,*

1. *for all $j < 0$ one has $S_j \cap B_2 = \emptyset$;*

2. *for all $j \geq 0$ one has $|\Gamma_{B_1}(S_j \cap A_1)| \geq (1/\alpha_j)|S_j \cap A_1|$*

3. *for all edges $e = (u, v) \in (A_1 \times B_2) \cap E_1$ one has $u \in S_j$ for some $j \geq 0$ and $v \in T_i$ for some i, $0 \leq i \leq j$.*

Proof: We start by showing part (1) of the lemma. By the choice of the cut $(A_1 \cup B_1, A_2 \cup B_2)$ all of A_2 can be matched to $Q \setminus B_1$ in $G_1' \cup G_2$. Let $T^* = \Gamma_{G_1'}(S_j \cap B_2)$. One has $|T^*| \geq (1/\alpha_j)|S_j \cap B_2|$. Hence, since vertices only arrive on the P side, one has $|\Gamma_{G_1' \cup G_2}(T^*) \setminus B_1| \leq \alpha_j |T^*| < |T^*|$, which contradicts the choice of the cut $(A_1 \cup B_1, A_2 \cup B_2)$.

Part (2) follows directly by Lemma 34 together with the assumption that $(A_1 \times B_2) \cap E = \emptyset$. Now (3) follows from (1) together with the fact that edges $e \in E_1 \setminus E_1'$ that have one endpoint in $T_i, i \geq 0$ can only go to S_j for some $j \geq 0$ by construction of G_1'. ■

We now prove the main theorem of this section:

Theorem 43 *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 $G_1 \cup G_2$.*

Proof: Let $(S_j, T_j), j = -\infty, \ldots, +\infty$ be the pairs from the definition of G'. Consider a saturating cut $(A_1 \cup B_1, A_2 \cup B_2)$ in $G_1' \cup G_2$. Recall that $A_1, A_2 \subseteq P_1 \cup P_2, B_1, B_2 \subseteq Q$, $(A_1 \times B_2) \cap (E_1' \cup E_2) = \emptyset$, $ALG = |B_1| + |A_2|$.

Let $S := \bigcup_{j \geq 0} S_j$. Choose a maximum matching M in $G_1 \cup G_2$ such that M matches all of S, as guaranteed by Lemma 41. Define

$$K_j = \{v \in \Gamma_{G_1'}(S_j) \cap B_2 : M(v) \notin S\}$$
$$K_j^* = \{v \in \Gamma_{G_1'}(S_j) \cap B_1 : M(v) \notin S\}$$

By Lemma 42 there are no edges in G_1 from $T_j, j < 0$ to B_2. This implies that

$$((A_1 \setminus S) \times B_2) \cap (E_1 \cup E_2) = \emptyset. \tag{20}$$

This allows us to obtain the following bound on the size of the matching M, which we denote by OPT. It follows from 20 that a matching edge that has an endpoint in $A_1 \setminus S$ necessarily has the other endpoint either in K_j^* for some j or in $B_1 \setminus \Gamma_{G_1'}(S)$. Hence, we have

$$OPT \leq |S| + \sum_{j \geq 0}(|K_j| + |K_j^*|) + (|B_1 \setminus \Gamma_{G_1'}(S)| + |A_2 \setminus (S \cup \bigcup_{j \geq 0} M(K_j))|). \tag{21}$$

Indeed, if an edge $\epsilon \in M$ has an endpoint in S, it is counted by the first term. Otherwise if e has an endpoint in $\Gamma_{G_1'}(S_j) \cap B_2$ for some j, it is counted in K_j; if e has an endpoint in $\Gamma_{G_1'}(S_j) \cap B_1$ for some j, it is counted in K_j^*. Finally, if e satisfies none of the above conditions, it must have one endpoint in either $B_1 \setminus \Gamma_{G_1'}(S)$ or $A_2 \setminus (S \cup \bigcup_{j \geq 0} M(K_j))$ by 20. Note that an edge $e \in M$ may satisfy more than one of these conditions, and hence we are only getting an upper bound on OPT.

By definition of the cut $(A_1 \cup B_1, A_2 \cup B_2)$ we also have

$$\begin{aligned} ALG =& |B_1| + |A_2| = |S \cap A_2| + \sum_{j \geq 0} |M(K_j)| + |\Gamma_{G_1'}(S) \cap B_1| \\ &+ (|B_1 \setminus \Gamma_{G_1'}(S)| + |A_2 \setminus (S \cup \bigcup_{j \geq 0} M(K_j))|), \end{aligned} \tag{22}$$

where we use the fact that $M(K_j) \subseteq A_2 \setminus S$ by definition of K_j together with 20. Thus, since $|M(K_j)| = |K_j|$, it is sufficient to show that

$$|S \cap A_2| + \sum_{j \geq 0} |K_j| + |\Gamma_{G_1'}(S) \cap B_1| \geq (3/4)(|S| + \sum_{j \geq 0} |K_j| + |K_j^*|)$$

Let

$$ALG' = |S \cap A_2| + \sum_{j \geq 0} |K_j| + |\Gamma_{G_1'}(S) \cap B_1|$$

$$OPT' = |S| + \sum_{j \geq 0} |K_j| + |K_j^*|.$$

Let $x_j := |S_j|, z_j = |S_j \cap A_1|, w_j := |\Gamma_{G_1'}(S_j \cap A_1)|, r_j^* := |K_j^*|, r_j := |K_j|$.

We will derive relations between these variables using the properties of the matching skeleton. By construction of G_1' we have

$$(S_i \times T_j) \cap E_1 = \emptyset, \forall i < j. \tag{23}$$

Define *canonical cuts* (U_k, W_k) as

$$U_k = \bigcup_{j=0}^{k} S_j \subseteq P_1, W_k = \bigcup_{j=0}^{k} T_j \subseteq Q. \tag{24}$$

By (23) we have that $(U_k \times (Q \setminus W_k)) \cap (E_1 \cup E_2) = \emptyset$.

Since M matches all of S, we have using the fact that canonical cuts are empty that for each $k \geq 0$

$$|U_k| \leq |W_k| - \sum_{j=0}^{k} (|K_j| + |K_j^*|).$$

Since $|T_j| = \alpha_j |S_j|$ by definition of G_1' and since T_j are disjoint, this can be equivalently stated in terms of the new variables as

$$\sum_{j=0}^{k} ((1/\alpha_j) x_j - r_j - r_j^*) \geq \sum_{j=0}^{k} x_j, \forall k \geq 0. \tag{25}$$

Thus, in terms of the new variables we have

$$OPT' = \sum_{j=0}^{\infty} x_j + \sum_{j=0}^{\infty} r_j + \sum_{j=0}^{\infty} r_j^*. \tag{26}$$

Similarly,

$$ALG' = \sum_{j=0}^{\infty}(x_j - z_j) + \sum_{j=0}^{\infty} w_j + \sum_{j=0}^{\infty} r_j \tag{27}$$

By Lemma 42, (3), we have $|\Gamma_{B_1}(S_j \cap A_1)| = w_j \geq (1/\alpha_j)z_j$.

Thus, putting (26), (27), (25) together, we have that it is sufficient to lower bound the solution of 28, obtaining a lower bound of P_1^* on the ratio ALG'/OPT', and hence on ALG/OPT.

$$\begin{aligned} P_1^* = \text{minimize } & \sum_{j=0}^{\infty}(x_j - z_j) + w_j + r_j \\ \text{s.t.} & \\ & \sum_{j=0}^{\infty}(x_j + r_j + r_j^*) \geq 1 \\ & \sum_{j=0}^{k}((1/\alpha_j)x_j - r_j - r_j^*) \geq \sum_{j=0}^{k} x_j, \forall k \\ & r_j^* \leq w_j \\ & w_j \geq (1/\alpha_j)z_j \\ & x, z, w, r, r^* \geq 0 \end{aligned} \tag{28}$$

We now transform 28 in two steps to obtain bounds $P_3^* \leq P_2^* \leq P_1^*$, and then show that $P_3^* \geq 3/4$.

First note that at the optimum one has $r \equiv 0$ since decreasing r and scaling all variables appropriately does not violate any constraints and only improves the solution. Next, we show that at the optimum, the third constraint is necessarily tight for all k. Otherwise let k be such that the constraint is not tight and let k^* be the smallest such that $k^* > k$ and $r_{k^*} > 0$.

Let

$$\begin{aligned} x' &= x \\ r^{*\prime} &= r^* + \delta e_k - \delta e_{k^*} \\ w' &= w + \delta e_k - \delta e_{k^*} \\ z' &= z + \alpha_k e_k - \alpha_{k^*} e_{k^*}. \end{aligned}$$

Note that $x', r^{*\prime}, w', z'$ form a feasible solution if $\delta > 0$ is sufficiently small. Finally,

$$\sum_{j=0}^{\infty}(x'_j - z'_j) + w'_j = \left(\sum_{j=0}^{\infty}(x_j - z_j) + w_j\right) + \delta(-\alpha_k + \alpha_{k^*}) < \sum_{j=0}^{\infty}(x_j - z_j) + w_j.$$

Also, for fixed r^*, x one can maximize z_j pointwise, so $r_j^* = (1/\alpha_j)z_j$ for all j.

Thus, we have $P_2^* \leq P_1^*$, where

$$
\begin{aligned}
P_2^* = \text{minimize } & 1 - \sum_{j=0}^{\infty} z_j \\
\text{s.t.} & \\
& \sum_{j=0}^{\infty} (x_j + (1/\alpha_j) z_j) = 1 \\
& \sum_{j=0}^{k} (1/\alpha_j - 1) x_j \geq \sum_{j=0}^{k} (1/\alpha_j) z_j, \forall k \\
& x, z \geq 0
\end{aligned}
\tag{29}
$$

Finally, we show that constraints in line 2 are necessarily tight at the optimum. Otherwise let k^* be the smallest such that constraint 2 is slack. Note that we necessarily have $x_{k^*} > 0$. Let

$$x' = x - \delta e_{k^*} + \delta e_{k^*+1} \frac{1/\alpha_{k^*} - 1}{1/\alpha_{k^*+1} - 1},$$

which is feasible for sufficiently small $\delta > 0$ and makes constraint 2 satisfied for all k. Let

$$\gamma = \sum_{j=0}^{\infty} (x_j + \alpha_j z_j) = 1 - \delta \left(1 - \frac{1/\alpha_{k^*} - 1}{1/\alpha_{k^*+1} - 1}\right) = 1 - \delta \frac{1/\alpha_{k^*+1} - 1/\alpha_{k^*}}{1/\alpha_{k^*+1} - 1} < 1.$$

Now $x'' = x'/\gamma, z'' = z/\gamma$ are feasible solutions that improve the objective function.

Thus, we have $x_j = z_j/(1 - \alpha_j)$ for all $j > 0$ (note that $x_0 = 0$ at the optimum for the same reason as $r \equiv 0$). Thus, we get $P_3^* \leq P_2^*$, where

$$
\begin{aligned}
P_3^* = \text{minimize } & 1 - \sum_{j=1}^{\infty} z_j \\
\text{s.t.} & \\
& \sum_{j=1}^{\infty} (1/(1 - \alpha_j) + 1/\alpha_j) z_j = 1 \\
& z \geq 0
\end{aligned}
\tag{30}
$$

In order to lower bound P_3^*, it is sufficient to minimize $f(\alpha) = 1/(1 - \alpha) + 1/\alpha$ over all $\alpha \in (0, 1]$. One has $f'(\alpha) = 1/(1 - \alpha)^2 - 1/\alpha^2$, $f'(1/2) = 0$ and $f''(\alpha) = 2/(1 - \alpha)^3 + 2/(1 - \alpha)^3 > 0$. Hence, the unique minimum is attained at $\alpha = 1/2$.

Thus, we have $z_j = 1/4$ for $\alpha_j = 1/2$ and zero otherwise. The objective value is $3/4$, proving that $3/4 \leq P_3^* \leq P_2^* \leq P_1^*$, and hence $ALG/OPT \geq 3/4$. ■

F 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 44 *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

$$G'_1 = \text{SPARSIFY}^*(G_1), \text{ and } G'_i = \text{SPARSIFY}^*(G'_{i-1} \cup G_i). \tag{31}$$

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 45 *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 46 *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

$$|\Gamma_{G'_\tau}(R^*)| \geq |\Gamma_{G'_{\tau+1}}(R^*)| \geq (1/\alpha^*)|R^*| > (1/\alpha)|R^*|. \tag{32}$$

Since $|\Gamma_{G'_\tau}(S_\alpha(\tau))| = (1/\alpha)|S_\alpha(\tau)|$, (32) 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 46 for all nonempty $S_{\alpha,\beta}$. We will need the following

Lemma 47 *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 46, 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 44. ■

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

$$\sum_{\beta \in [\alpha,1]} \sum_{\delta \in (0,\alpha-\Delta]} s_{\beta,\delta}(\tau) \leq (\alpha - \Delta) \sum_{\beta \in [\alpha,1]} t_\beta(\tau). \tag{33}$$

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 46 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

$$\sum_{\beta \in [\alpha,1]} t_\beta(\tau+1) \geq \sum_{\beta \in [\alpha,1]} t_\beta(\tau) + \sum_{\gamma \in (0,\alpha-\Delta]} \frac{1}{\gamma} r_\gamma(\tau) \geq \sum_{\beta \in [\alpha,1]} t_\beta(\tau) + \frac{1}{\alpha - \Delta} \sum_{\gamma \in (0,\alpha-\Delta]} r_\gamma(\tau). \tag{34}$$

Furthermore, by Lemma 46

$$\sum_{\beta \in [\alpha,1]} \sum_{\delta \in (0,\alpha-\Delta]} s_{\beta,\delta}(\tau+1) = \sum_{\beta \in [\alpha,1]} \sum_{\delta \in (0,\alpha-\Delta]} s_{\beta,\delta}(\tau) + \sum_{\gamma \in (0,\alpha-\Delta]} r_\gamma(\tau) \tag{35}$$

Since by inductive hypothesis

$$\sum_{\beta\in[\alpha,1]} t_\beta(\tau) \geq \frac{1}{\alpha-\Delta} \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} s_{\beta,\delta}(\tau). \tag{36}$$

we have by combining (34), (35) and (36)

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

■

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 41. 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 47). 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 47.

Lemma 49 *For all* $\lambda \leq 1$

$$\sum_{\alpha\in[\lambda,1]} t_\alpha \geq \sum_{\alpha\in[\lambda,1],\beta\in[\lambda,1]} s_{\alpha,\beta} + \sum_{\alpha\in[\lambda,1]} r_\alpha + \gamma. \tag{37}$$

Proof: Follows from Lemma 47. ■

We also have

$$\sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,1]} s_{\beta,\delta} = \sum_{\beta\in[\alpha,1]} \beta t_\beta \tag{38}$$

for all $\alpha \in (0,1]$.

By Lemma 48 and Lemma 49, 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

$$\begin{aligned}
&\forall \alpha \in (0,1] : \sum_{\beta\in[\alpha,1]} t_\beta \geq \gamma + \sum_{\beta\in[\alpha,1]} \sum_{\delta\in[\alpha,1]} s_{\beta,\delta} + \sum_{\beta\in[\alpha,1]} r_\beta. \\
&\forall \alpha \in (0,1] : \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} s_{\beta,\delta} \leq (\alpha-\Delta) \sum_{\beta\in[\alpha,1]} t_\beta \\
&\forall \alpha \in (0,1] \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,1]} s_{\beta,\delta} = \sum_{\beta\in[\alpha,1]} \beta t_\beta.
\end{aligned} \tag{39}$$

We start by simplifying (39). 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 (40), obtaining a lower bound of $\frac{P_1^*}{P_1^*+1}$ on the ratio ALG/OPT.

$$
\begin{aligned}
P_1^* = \text{minimize} \quad & \sum_{\alpha\in(0,1)} (t_\alpha - r_\alpha) \\
\text{s.t.} \quad & \\
\forall \alpha \in (0,1] : \quad & \sum_{\beta\in[\alpha,1]} t_\beta \geq 1 + \sum_{\beta\in[\alpha,1]} \sum_{\delta\in[\alpha,1]} s_{\beta,\delta} + \sum_{\beta\in[\alpha,1]} r_\beta. \\
\forall \alpha \in (0,1] : \quad & \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,\alpha-\Delta]} s_{\beta,\delta} \leq (\alpha - \Delta) \sum_{\beta\in[\alpha,1]} t_\beta \\
\forall \alpha \in (0,1] \quad & \sum_{\beta\in[\alpha,1]} \sum_{\delta\in(0,1]} s_{\beta,\delta} = \sum_{\beta\in[\alpha,1]} \beta t_\beta \\
& t_\alpha, s_{\alpha,\beta} \geq 0.
\end{aligned}
\tag{40}
$$

Combining constraints 2 and 3 of (40), we get

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

Thus, it is sufficient to lower bound the optimum of

$$
\begin{aligned}
P_2^* = \text{minimize} \quad & \sum_{\alpha\in(0,1)} (t_\alpha - r_\alpha) \\
\text{s.t.} \quad & \\
\forall \alpha \in (0,1] : \quad & \sum_{\beta\geq\alpha} (1 - \beta + \alpha - \Delta) t_\beta \geq 1 + \sum_{\beta\in[\alpha,1)} r_\alpha. \\
& t_\alpha \geq 0.
\end{aligned}
\tag{41}
$$

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:

$$
\begin{aligned}
P_3^* = \text{minimize} \quad & \sum_{\alpha \in (0,1)} t_\alpha \\
& \text{s.t.} \\
\forall \alpha \in (0,1] : & \sum_{\beta \geq \alpha} (1 - \beta + \alpha - \Delta) t_\beta \geq 1. \\
& t_\alpha \geq 0.
\end{aligned}
\tag{42}
$$

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

$$\sum_{\beta \geq \alpha} (1 + (\alpha - \beta - \Delta)) t_\beta = 1, \forall \alpha \in [0,1]. \tag{43}$$

Subtracting (43) for $\alpha + \Delta$ from (43) for α, we get

$$
\begin{aligned}
\sum_{\beta \geq \alpha} (1 + (\alpha - \beta - \Delta)) t_\beta - \sum_{\beta \geq \alpha + \Delta} & (1 + (\alpha + \Delta - \beta - \Delta) t_\beta \\
&= t_\alpha - \Delta \sum_{\beta \geq \alpha} t_\beta = 0.
\end{aligned}
\tag{44}
$$

In other words,

$$t_\alpha = \Delta \sum_{\beta \geq \alpha} t_\beta, t_1 \geq 1. \tag{45}$$

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,

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

Hence, one has

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

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

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

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

Theorem 50 *There exists a deterministic $O(n)$ space 1-pass streaming algorithm for approximating the maximum matching in bipartite graphs in the vertex arrival model.*

Proof: Run the algorithm given in (31), 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 $O(n)$. ■

G 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 H 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 H to prove that our bound on $CC_v(\epsilon, n), \epsilon < 1/4$ is tight. Specifically, we show the following results:

Lemma 51 *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 52 *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 [17]), 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.

G.1 Balanced graphs

The following lemma is an adaptation of Theorem 16 in [7] (see also [17]), where *bipartite* $1/3$-Ruzsa-Szemerédi graphs with a superlinear number of edges are constructed. The main idea of the construction, i.e. the 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 the argument here.

Lemma 53 *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.*

Proof: Let $X = Y = [m^2]^m$ for some integer $m > 0$. We will refer to vertices in X and Y as points in $[m^2]^m$. Matchings M_T will be indexed by subsets $T \subseteq [m]$.

Fix $T \subseteq [m]$. Let $L_s = \{x : \sum_{i \in T} x_i = s\}$. Define red, white and blue strips as follows. Choose $w = 2(1 + 2/\epsilon)(\epsilon m/6)$and define

$$R_k = \bigcup_{s=kw}^{kw+(2/\epsilon)(\epsilon m/6)-1} L_s$$

$$W_k = \bigcup_{s=kw+(2/\epsilon)(\epsilon m/6)}^{kw+(1+(2/\epsilon))(\epsilon m/6)-1} L_s$$

$$B_k = \bigcup_{s=kw+(1+2/\epsilon)(\epsilon m/6)}^{kw+(1+4/\epsilon)(\epsilon m/6)-1} L_s$$

$$W'_k = \bigcup_{s=kw+(1+4/\epsilon)(\epsilon m/6)}^{(k+1)w-1} L_s$$

Finally, define $B = \bigcup_k B_k, R = \bigcup_k R_k, W' = \bigcup_k W'_k, W = \bigcup_k W_k$.

For $T \subseteq [m]$ let 1_T denote the characteristic vector of T. The matching M_T is defined as follows. If a blue point $b \in B^X$ has all coordinates greater than $(2/\epsilon + 1)$, match it to the point $r = b - (2/\epsilon + 1) \cdot 1_T$ in R^Y. Note that $r \in R^Y$ by the definition of B and R.

Following [7], we first note that

Lemma 54 $|M_T| \geq (1/2 - \epsilon)n - o(n)$

Proof: The only points of B that are not matched by M_T are those in the set

$$S = \{x : \exists j \in T, x_j < (2/\epsilon + 1)v_j\}.$$

However, $|S| \leq \frac{(m/6)(2/\epsilon+1)}{m^2}|X| = \frac{(2/\epsilon+1)}{6m}|X| = o(|X|)$. Hence, we have that $|B| = (1 \pm o(1))|R|$. Similarly, we have that $|W| \leq (\epsilon/(1+\epsilon) \pm o(1))|B|$ ∎

Now let T_1, T_2 be two sets in $[m]$ of size $(\epsilon/6)m$ such that $|T_1 \cap T_2| \leq (5/2)(\epsilon/6)^2 m$. We show that no edge of M_{T_1} is induced by M_{T_2}. Let b be matched to r by T_1, i.e. $b - r = (2/\epsilon + 1)1_{T_1}$. If the edge (b, r) is induced by M_{T_2}, then one of b, r is colored blue and the other is colored red in the coloring induced by T_2. In particular, b and r are separated by a white strip. Thus,

$$\left| \sum_{i \in T_2} b_i - \sum_{i \in T_2} r_i \right| \geq (\epsilon/6)m. \tag{46}$$

On the other hand,

$$\left|\sum_{i\in T_2} b_i - \sum_{i\in T_2} r_i\right| = \left|\sum_{i\in T_2}(b-r)_i\right| = \left|\sum_{i\in T_2}((2/\epsilon+1)1_{T_1})_i\right| \tag{47}$$
$$= (2/\epsilon+1)|T_1\cap T_2| < (2/\epsilon+1)(5/2)(\epsilon/6)^2 m = (5/6)(1+\epsilon/12)(\epsilon/6)m,$$

a contradiction with (46) for any $\epsilon \le 1/2$.

Now it suffices to exhibit a large family $\mathcal{F}$ of subsets of $[m]$ of size $(\epsilon/6)m$ with intersection at most $(5/2)(\epsilon/6)^2$. Following [7], we obtain such a family from an error-correcting code with weight $w = (\epsilon/6)m$ and Hamming distance at least $d = 2(\epsilon/6) - (5/2)(\epsilon/6)^2$. The Gilbert-Varshamov bound yields [15], for $d \le \frac{2w(m-w)}{m}$, a family $\mathcal{F}$ such that

$$\frac{1}{m}\log|\mathcal{F}| \ge H\left(\frac{w}{m}\right) - \frac{w}{m}H\left(\frac{d}{2w}\right) - \left(1-\frac{w}{m}\right)H\left(\frac{d}{2(m-w)}\right) - o(1)$$

Letting $\delta = \epsilon/3$ and $\gamma = 5/4$ for convenience, we have that $w/m = \delta$ and $d/m = 2\delta(1-\gamma\delta)$. This yields

$$\frac{1}{m}\log|\mathcal{F}| \ge H(\delta) - \delta H(1-\gamma\delta) - (1-\delta)H\left(\frac{\delta-\gamma\delta^2}{1-\delta}\right) - o(1)$$

Using $H(x) = H(1-x)$ and strict convexity of $H(x)$, we get

$$\delta H(1-\gamma\delta) + (1-\delta)H\left(\frac{\delta-\gamma\delta^2}{1-\delta}\right) \le c(\delta,\gamma) + H\left(\gamma\delta^2 + (1-\delta)\left(\frac{\delta-\gamma\delta^2}{1-\delta}\right)\right) = c(\delta,\gamma) + H(\delta)$$

where $c(\delta,\gamma) > 0$ whenever $\gamma \ne 1$.

Hence, setting $\gamma = 5/4$ and $\delta = \epsilon/6$ yields a family of codes with $\frac{1}{m}\log|\mathcal{F}| \ge c(\epsilon/6, 5/4) - o(1)$.

Thus, we have constructed a bipartite graph $G = (X, Y, E)$ such that $E = \bigcup_{T\in\mathcal{F}} M_T$ is a union of induced matchings of size $1/2 - \epsilon - o(1)$. The number of nodes in the graph is m^{2m} and the number of matchings is $|\mathcal{F}| = 2^{(c(\epsilon/3,5/4)-o(1))m} = 2^{\Omega(m)}$. Thus, we get a graph on $n = m^{2m}$ nodes that is a union of $2^{\Omega(m)} = n^{\Omega_\epsilon(1/\log\log n)}$ induced matchings of size $1/2 - \epsilon$. ■

G.2 Lop-sided graphs

We now extend this construction to lop-sided graphs, which will be important for showing optimality of our bound on $CC_v(\epsilon, n)$.

Lemma 55 *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 $(1/2 - O(\delta))|X|$.*
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 $Y \setminus (M_j \cap Y)$.*

Proof: Let $X' = Y = [m^2]^m$ for some integer $m > 0$. Let X be a random subset of X' that contains each element of X' with probability $1/2$. We will refer to vertices in X and Y as points in $[m^2]^m$. The matchings M_T will be indexed by subsets $T \subseteq [m]$.

Choose $w = 2C(1+2/\delta)p$ for p and a constant $C > 0$ to be specified later. Fix $T \subseteq [m]$. Let $L_s = \{x : \sum_{i\in T} x_i = s + (w/2)\cdot Q_T\}$, where Q_T is a Bernoulli $0/1$ random variable with on probability $1/2$. Define red, white and blue strips as follows. Define

$$R_k = \bigcup_{s=kw}^{kw+C(2/\delta)p-1} L_s$$
$$W_k = \bigcup_{s=kw+C(2/\delta)p}^{kw+C(1+(2/\delta))p-1} L_s$$
$$B_k = \bigcup_{s=kw+C(1+2/\delta)p}^{kw+C(1+4/\delta)p-1} L_s$$
$$W'_k = \bigcup_{s=kw+C(1+4/\delta)p}^{(k+1)w-1} L_s$$

Define $B = \bigcup_k B_k, R = \bigcup_k R_k, W' = \bigcup_k W'_k, W = \bigcup_k W_k$.

Here we are assuming that $\delta \in (0,1)$ is such that $2/\delta$ is an integer.

Fix k. For two vertices $u, v \in B_k$ we say that $u \sim v$ if $u - v = \lambda(1+2/\delta)\cdot 1_T$ for some λ (note that since $u, v \in B_k$, we have $\lambda \in [-C, C]$). We write $S_v \subseteq Y$ to denote the equivalence class of v. Note that $|S_v| \geq C/2$ for all v. Also, let

$$T_v = \{u \in X : u = w - C(1+2/\delta)\cdot 1_T, w \in S_v\}.$$

Note that for any $v \in B_k$ one has $T_v \subseteq R_k$. Note that T_v is a random set (determined by the random choice of $X \subset X'$).

We now define a 2-matching from (a subset of) T_v to S_v. First note that $\mathbf{E}[|T_v|] = \frac{1}{2}|S_v|$. Furthermore, since X is obtained from X by independent sampling, the events $\{v \in X\}$ are independent conditional on the value of Q_T. Thus, standard concentration inequalities apply (see, e.g. [10]) and we get

$$\mathbf{Pr}\left[|T_v| \notin (1\pm\delta)\frac{1}{2}|S_v|\right] \leq e^{-\delta^2(1/2)|S_v|/4} = e^{-\delta^2 C/16} \leq \delta/4$$

for $C > 16\ln(4/\delta)/\delta^2$. We now classify points $v \in B_k$ as good or bad depending on the how close $|T_v|$ is to its expectation. In particular, mark v *bad* if $|T_v| \notin (1\pm\delta)\frac{1}{2}|S_v|$ and *good* otherwise. If v is good, let T'_v denote an arbitrary subset of T_v of cardinality $(1-\delta)\frac{1}{2}|S_v|$. Similarly, let S'_v denote an arbitrary subset of S_v of cardinality $(1-\delta)|S_v|$, so that $|T'_v| = \frac{1}{2}|S'_v|$. Next, choose an arbitrary 2-matching from T'_v to S'_v. Note that all matched edges are of the form (b, r), where $r = b - \lambda(1+2/\delta)\cdot 1_T$ for some $\lambda \in (0, 2C]$. This completes the definition of the 2-matching M_T for a fixed set T.

We now argue that there cannot be too many bad classes in a fixed set T. Note that there are $\Omega(m^{2m})$ equivalence classes (since they have constant size by construction). For a vertex v denote the event that v's equivalence class is bad by $\mathcal{E}_v$. Then, conditional on the value of Q_T, these events are independent for non-equivalent v's. Hence, by Chernoff bounds the probability that the number of bad classes exceeds its expectation by more than a factor of 4 is at most $e^{-\Omega(m^{2m})}$. We use the collection $\mathcal{F}$ constructed in the proof of Lemma 53, and a union bound over $2^{O(m)}$ sets T shows that there will be no more than a δ fraction of bad classes in any of sets T with high probability.

We will also need a bound on the maximum degree of vertices in X and Y. First note that the definition of the set of levels L_s and the random variable Q_T amounts to flipping the role of the sets R_k and B_k

independently with probability $1/2$. Thus, for a fixed T, every vertex except for those in $W \cup W'$, of which there is only an $O(\delta)$ fraction, takes part in the matching with probability $1/2$. Thus, the expected degree of a fixed vertex $v \in Y$ is at most $|\mathcal{F}|/2$, where $\mathcal{F}$ is the collection of almost orthogonal vectors that we use. Since Q_T are independent for each T, Chernoff bounds imply that the degree of any vertex v in Y in the graph that we construct is at most $(1+\delta)|\mathcal{F}|/2$ with probability at least $1 - e^{-\Omega(|\mathcal{F}|)} = 1 - e^{-\Omega(2^{\Omega(m)})}$. In particular, a union bound over all $v \in Y$, of which there are m^{2m}, shows that the degree cannot be larger than $(1+\delta)|\mathcal{F}|/2$ for any v with high probability. Finally, we also note that the average degree is at least $(1 - O(\delta))|\mathcal{F}|/2$ by construction. A similar argument shows that the maximum degree of a vertex in X does not exceed $(1+\delta)|\mathcal{F}|$ with high probability, and the average degree is at least $(1 - O(\delta))|\mathcal{F}|$.

Essentially the same argument as in Lemma 53 together with the fact that for good sets T we have a 2-matching of at least $(1 - O(\delta))|R|$ nodes by the argument above shows that the size of the matching is at least $(\frac{1}{2} - O(\delta))|X|$.

Now let T_1, T_2 be two sets in $[m]$ of size $p = (\delta/(8C))m$ such that $|T_1 \cap T_2| \leq (5/2)(\delta/(8C))^2 m$. We show that no edge of M_{T_1} is induced by M_{T_2}. Let b be matched to r by T_1, i.e. $b - r = j(2/\delta + 1)1_{T_1}$ for some $j \in (0, 2C]$. If the edge (b, r) is induced by M_{T_2}, then one of b, r is colored blue and the other is colored red in the coloring induced by T_2. In particular, b and r are separated by a white strip. Thus,

$$\left| \sum_{i \in T_2} b_i - \sum_{i \in T_2} r_i \right| \geq (\delta/(8C))m. \tag{48}$$

On the other hand,

$$\begin{aligned} &\left| \sum_{i \in T_2} b_i - \sum_{i \in T_2} r_i \right| = \left| \sum_{i \in T_2} (b - r)_i \right| \\ &= \left| \sum_{i \in T_2} (j(2/\delta + 1)1_{T_1})_i \right| \\ &\leq 2C(2/\delta + 1)|T_1 \cap T_2| < C(2/\delta + 1)(5/2)(\delta/(8C))^2 \\ &\leq (5/6)(1 + \delta/12)(\delta/8C)m, \end{aligned} \tag{49}$$

a contradiction with (48) for any $\delta \leq 1/2$. This completes the proof of (1).

It remains to show (2). Consider a fixed matching M_T. Let $R_T = X \cap M_T \subseteq R^X, B_T = Y \cap M_T \subseteq B^Y$, where we use the notation B^X, B^Y, R^X, R^Y to denote the set of blue and red points in X and Y respectively. For a vertex $u \in X \cup Y$, denote by $\Gamma(u)$ its neighbors in G. Let

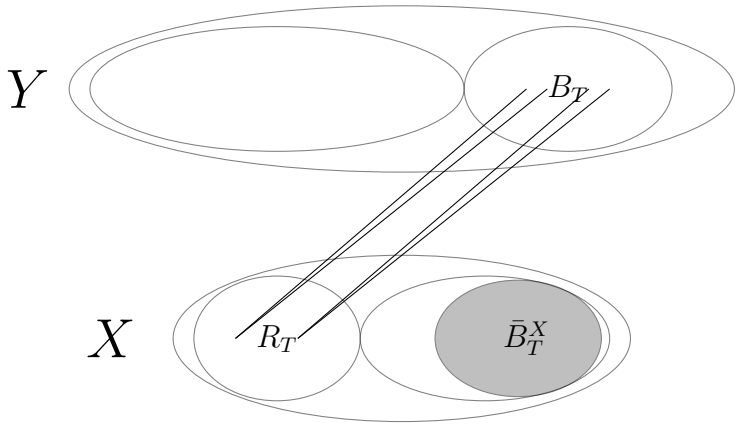
$$\bar{B}_T = \bigcup_k \bigcup_{s=kw+C(2+2/\delta)p}^{kw+C(4/\delta)p-1} L_s.$$

Note that $\bar{B}_T \subset B_T$ can be viewed as the 'interior' of B_T. We write $\bar{B}_T^X$ and $\bar{B}_T^Y$ to denote the projection of $\bar{B}$ onto X and Y respectively.

We first show that for all $x \in \bar{B}_T^X$ one has $\Gamma(x) \subseteq B_T \subseteq Y$. Since $\bar{B}_T^X$ is not matched by T, it suffices to consider edges of $M_{T'}, T' \neq T$. But any such edge has the form (x, y), where $x = y \pm \lambda(2/\delta + 1) \cdot 1_{T'}$, so by the argument above one has

$$\left| \sum_{i \in T_2} x_i - \sum_{i \in T_2} y_i \right| < (\delta/(8C))m = p, \tag{50}$$

Figure 4: A 2-matching M_T



so $y \in B_T^Y$.

Thus, for each $x \in \bar{B}_T^X$ one has $\Gamma(x) \subseteq B_T^Y$. We can now exhibit the required fractional matching. Include edge $(x,y), r \in B_T^X, y \in \bar{B}_T^Y$ with weight $\frac{1}{(1+O(\delta))|\mathcal{F}|}$, and include all edges of the 2-matching M_T with weight $1/2$. Since the maximum degree of a node in B_T is at most $(1+\delta)|\mathcal{F}|/2$, and the maximum degree of a node in $\bar{B}_T$ is at most $(1+O(\delta))|\mathcal{F}|$, this assignment yields a feasible fractional matching. Recall that by construction, the average degree in X is at least $(1-O(\delta))|\mathcal{F}|/2$, hence the size of the fractional matching is at least $(1-O(\delta))|X|$.

By the integrality of the matching polytope, the fractional matching can be rounded to produce an integral matching of size at least $(1-O(\delta))|X|$, as required. Note that since we proved that for each $x \in \bar{B}_T^X$ one has $\Gamma(y) \subseteq B_T^Y$, the fractional matching that we constructed avoids $Y \setminus B_T = Y \setminus (Y \cap M_T)$, and hence so does the integral matching. This completes the proof of (2).

Finally, we note that the number of edges in the graph is given by $n^{1+\Omega_\delta(1/\log\log n)}$, as before. ■

We note that the same techniques can be used to prove the following more general

Lemma 56 *For any fixed constants $\epsilon, \gamma > 0$ and an arbitrarily small constant $\delta > 0$ there exists a family of bipartite Ruzsa-Szemerédi graphs $G = (X, Y, E)$ with $|X| = n$, $|Y| = n/\epsilon$ such that*

1. *the edge set E is a union of $n^{\Omega_{\epsilon,\delta,\gamma}(1/\log\log n)}$ induced $\frac{1-\gamma}{\epsilon\gamma}$-matchings $M_1, \ldots, M_k$ of size at least $(\gamma-\delta)|X|$.*

2. *for every $j \in [1:k]$ the graph G contains a matching M_j^* of size at least $(1-O(\delta))|X|$ that avoids $Y \setminus (M_j \cap Y)$.*

H 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 53) 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 D 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 52, 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 E, 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.

H.1 Edge arrivals

Lemma 57 *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 $1 \leq i \leq k$, 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 58 *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 $|F| \geq 2^{(\gamma - o(1))U_I(\epsilon,n)}$, $|I| = o(k)$.*

Proof: Let $|I| = k_1$. Then the number of graphs that can be in F is bounded by

$$\binom{(\epsilon - \epsilon')n}{\delta n}^{k_1} \binom{\epsilon n}{\delta n}^{k-k_1} = \left(2^{-\Omega(\epsilon' n)} \binom{\epsilon n}{\delta n}\right)^{k_1} \binom{\epsilon n}{\delta n}^{k-k_1} = 2^{-\Omega(k_1(\epsilon' n))} \binom{\epsilon n}{\delta n}^{k}.$$

It then follows that if $k_1 = \Omega(k)$, we have $|F| \leq 2^{(\gamma - \Omega(1))U_I(\epsilon,n)}$, contradicting our assumption on the size of F. ■

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 58 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 59 *For any $\delta > 0$, $CC(2/3 + \delta, n) = n^{1+\Omega_\delta(1/\log\log n)}$.*

Proof: Follows by putting together Lemma 53 and Lemma 57. ■

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

Corollary 60 *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 53 and Lemma 57. ■

H.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 52. The bound implies that our upper bound from section E is tight. Moreover, the bound yields the first lower bound on the streaming complexity in the vertex arrival model.

Lemma 61 *For any constant $\delta > 0$, $CC_v^1(3/4 + \delta, n) = n^{1+\Omega_\delta(1/\log\log n)}$.*

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 55. 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 52, 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 57. 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 62 *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.*

I Matching covers versus 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 I.1.

Theorem 63 [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 of Theorem 63: 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 64), and then the full version (Theorem 65). The simple version is a corollary of the full version; the full version is proved in section I.1.

Theorem 64 [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 65 [Upper bound] *Assume $\epsilon n \geq 3$, and $0 < \delta < 1$. Then,*

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

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 64, we get the interesting special case, $L_C(n, \epsilon) = O(U_I(\epsilon/2, n)\log(1/\epsilon))$.

I.1 Proof of the Upper Bound

We will now prove Theorem 65. 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 66 *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.

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

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

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 67 *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 68 *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 65 is immediate from these two lemmas.

I.1.1 Proof of Lemma 67

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 67 in two parts.

Lemma 69 *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 70 *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

$$\begin{aligned}
&\prod_{e \in E \cap (A \times B)} (1 - \gamma x_e) \\
&\leq e^{-\gamma \sum_{e \in E \cap (A \times B)} x_e} \\
&\leq e^{-\gamma} \qquad \text{[From feasibility of the fractional solution].}
\end{aligned}$$

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 67.

I.1.2 Proof of Lemma 68

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 68.

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 71 $CC_+(\epsilon, n) \leq L_C(\epsilon, n)$.

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

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

■