CS 671: Graph Streaming Algorithms and Lower Bounds	**Rutgers: Fall 2020**

Lecture 2

September 15, 2020

Instructor: Sepehr Assadi — *Scribe: Sepehr Assadi*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

In this lecture, we will primarily focus on the following paper:

- "Ashish Goel, Michael Kapralov, Sanjeev Khanna, On the communication and streaming complexity of maximum bipartite matching. In SODA 2012."

1 The Maximum Bipartite Matching Problem

A *matching* in an undirected graph $G = (V, E)$ is a collection of edges that share no endpoints. A *perfect matching* is then a matching that matches all vertices (not every graph has a perfect matching). In the *maximum matching* problem, we are interested in finding a largest size matching in a given graph.

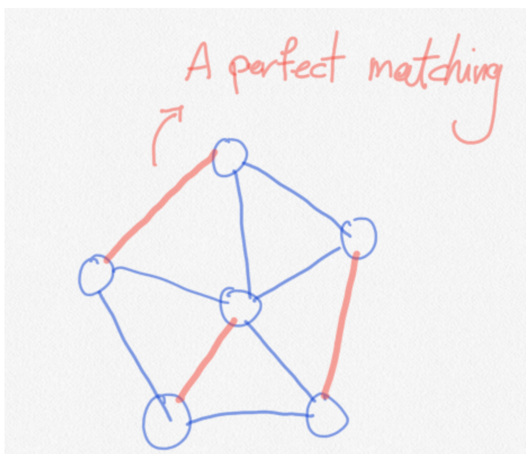

Figure 1: An example of a matching.

Matching problem is a fundamental problem in TCS and graph theory with a wide range of applications. A particularly important special case of this problem is *bipartite* matching on bipartite graphs $G = (L, R, E)$ which capture many natural applications of matching problem, for instance, matching items to buyers or clients to servers. Throughout this lecture (and this course in general), we mostly focus on the bipartite matching problem for simplicity although many of the ideas discussed will generalize to general graphs as well (and some do not).

The maximum bipartite matching problem has been studied extensively starting as early as the work of König [14] over a century ago, and continues to be an excellent testbed for development of fundamental algorithmic tools and ideas. This problem can be solved easily via a reduction to maximum flow in $O(mn)$ time[1] (this is a standard undergraduate algorithm exercise). There is also a simple elegant algorithm due to Hopcroft and Karp [12] for solving this problem in $O(m\sqrt{n})$ time (see also [18] for a very recent and exciting breakthrough on this problem that improves this runtime to $\widetilde{O}(m + n\sqrt{n})$[2]).

[1] Throughout the course, n and m denote the number of vertices and edges, respectively.
[2] This algorithm however is not really simple to put it mildly.

In this lecture, we study the bipartite matching problem in single-pass streams and one-way communication model; we will revisit this problem multiple times in this course in other settings as well. Before moving on from this section, let us set up some notation and also recall a fundamental result about bipartite matchings.

Notation. We use $\mu(G)$ to denote the size of a maximum matching of the graph G. Throughout, we will always assume $|L| = |R| = n/2$ (which can be obtained by a simple padding). For any subset of vertices S, $N(S)$ denotes the neighborset of S and $E(S)$ is the set of edges incident on S; when working with more than one graphs on the same vertices with subscript these with the corresponding graph to avoid confusion.

Hall's Marriage Theorem. A classical result in graph theory is Hall's marriage theorem of [11] that characterize bipartite graphs that admit a perfect matching.

Proposition 1 (Hall's theorem [11])**.** *A bipartite graph $G = (L, R, E)$ has a perfect matching iff:*

$$\forall S \subseteq L : |S| \leq |N(S)| \qquad \textit{and} \qquad \forall T \subseteq R : |T| \leq |N(T)|,$$

where $N(A)$ is the set of neighbors of the set A in G.

Proof. A simple way of proving this is to write the max-flow reduction for bipartite matching and apply the max-flow min-cut duality (although more elementary proofs by induction are also possible). $\square$

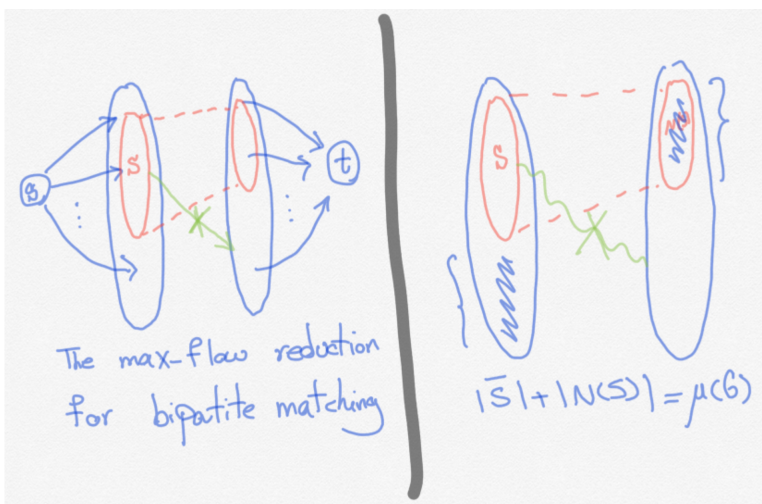

Figure 2: An illustration of Hall's theorem and its extension.

This theorem can be easily extended to characterize the size of a maximum matching in bipartite graphs.

Proposition 2 (Extended Hall's theorem (folklore))**.** *For any bipartite graph $G = (L, R, E)$,*

$$\mu(G) = \frac{n}{2} - \max_S(|S| - |N(S)|),$$

where S ranges over all subsets of L and R (separately)—any arg max *above is called a* witness set*.*

Proof. Let $k := \max_S(|S| - |N(S)|)$ and S be any witness set to it and by symmetry assume S belongs to L.

We first have that $\mu(G) \leq n/2 - k$: this is because at least k vertices of S cannot be matched as size of neighborhood of S is $|S| - k$. As such, even if all other vertices of L gets matched in G, size of the matching would be $|L \setminus S| + |S| - k = n/2 - k$.

Let us now prove that $\mu(G) \geq n/2 - k$. Consider the graph G' obtained by copying G, adding k new vertices to each side of the bipartition, and connecting those vertices to all original vertices of G in G'. We can verify that in this new graph, for any subset T of vertices in each bipartition (separately), $|T| \leq |N(T)|$ (as we "removed the deficit" in $|S| - |N(S)|$ for every set S in G). As such, by Hall's theorem in Proposition 1, G' has a perfect matching M' with size $n/2 + k$. As the new $2k$ vertices in G' can only match $2k$ edges in M', this means that there is a matching $M \subseteq M'$ that do not use any of the new vertices and has size $n/2 + k - 2k = n/2 - k$. Since $M \subseteq G$, we get that $\mu(G) \geq n/2 - k$. The two parts above imply that $\mu(G) = n/2 - k = \max_S(|S| - |N(S)|)$, finalizing the proof. $\square$

Remark. A similar-in-spirit but considerably more involved characterization for size of maximum matchings in general (not necessarily bipartite) graphs is the **Tutte-Berge formula**: It states that in any graph $G = (V, E)$, size of the maximum matching is:

$$\mu(G) = n + \frac{1}{2} \min_{U \subseteq V} \left(|U| - odd(V \setminus U) \right),$$

where $odd(V \setminus U)$ counts how many connected components of the subgraph $V \setminus U$ have an odd number of vertices. We will not work with this formula in this course as our focus is primarily on bipartite graphs (if you want to familiarize yourself a bit with this formula, use it to give a short proof of Proposition 2).

2 Bipartite Matching in Single-Pass Streams

Perfect Matchings and Exact Algorithms

Let us start by proving that no single-pass streaming algorithm can decide whether a graph has a perfect matching or not in $o(n^2)$. This lower bound was first proved by Feigenbaum et.al. [7].

Theorem 3. *Single-pass streaming algorithms for finding a perfect matching require $\Omega(n^2)$ space.*

Proof. The proof is by a simple reduction from the Index problem on the domain $\{0,1\}^N$ where $N = \Theta(n^2)$ where n is the number of vertices of the graph. The reduction is as follows:

(i) The players set the vertices of the graph to be $L := L_1 \cup L_2$ and $R := R_1 \cup R_2$, where $|L_1| = |R_1| = \sqrt{N}$ and $|L_2| = |R_2| = \sqrt{N} - 1$ (thus G has $n = \Theta(\sqrt{N})$ vertices). Define $\sigma : [N] \to [\sqrt{N}] \times [\sqrt{N}]$ to be any fixed bijective mapping, e.g., $\sigma(i) \to (i/\sqrt{N}, i \mod \sqrt{N})$.

(ii) Given $x \in \{0,1\}^N$, for all $x_i = 1$, Alice adds an edge between $\ell_p \in L_1$ and $r_q \in R_1$ where $\sigma(i) = (p, q)$.

(iii) Given $i \in [N]$ with $\sigma(i) = (p, q)$, Bob adds a perfect matching M_{21} between L_2 and $R_1 \setminus \{q\}$ and another one M_{12} between R_2 and $L_1 \setminus \{p\}$.

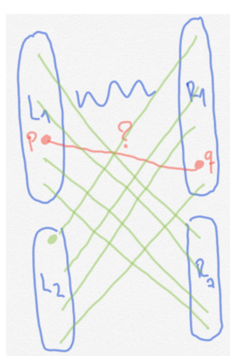

Figure 3: An illustration of the reduction in Theorem 3.

We claim that the resulting graph G has a perfect matching iff $Ind(x, i) = x_i = 1$ as follows:

(i) If $Ind(x, i) = x_i = 1$, there is a perfect matching in G consists of $M_{21} \cup M_{12} \cup \{(\ell_p, r_q)\}$ for $\ell_p \in L_1$, $r_q \in R_1$ and $(p, q) = \sigma(i)$.

(ii) If $Ind(x, i) = x_i = 0$, there is no perfect matching in G. Consider the set $S = L_2 \cup \{\ell_p\}$ for $(p, q) = \sigma(i)$; then $N(S) = R_1 \setminus \{r_q\}$ as vertices of L_2 have no edge to r_q by the choice of Bob and ℓ_p has no edge to r_q as $x_i = 0$. This mean $|S| > |N(S)|$ and G has no perfect matching by Hall's theorem (Proposition 1).

As such any one-way communication protocol for bipartite matching requires $\Omega(n^2)$ communication. Combining this with the connection to streaming algorithms implies the theorem. $\square$

Remark. The lower bound of Theorem 3 is the prototypical example of an $\Omega(n^2)$ lower bound based on Index, wherein Alice has a bipartite graph and input of Bob makes exactly one pairs of vertices "important" for solving the problem—thus, we need to figure out whether or not that edge existed in Alice's input. As an exercise, you can use this to also prove that any single-pass streaming algorithm for directed s-t reachability or undirected shortest path requires $\Omega(n^2)$ space.

Theorem 3 indicates the we cannot hope to achieve any non-trivial streaming algorithm for perfect matching or maximum matching (even determining their size and not necessarily their edges) in a single pass; all we can do is to basically use the trivial algorithm that stores the entire input graph. This lower bound was very recently extended to two pass algorithms in [3]; we also know that any semi-streaming algorithm for this problem requires $\Omega(\log n / \log\log n)$ passes [10]. We will study these lower bounds later in this course.

A Simple 2-Approximation

Let us now show that there is a very simple algorithm for obtaining a 2-approximation to matching in a single pass. The algorithm goes over the edges and greedily pick any edge that can be inserted to the currently maintained matching, and ignores the rest.

Algorithm 1. A single-pass semi-streaming algorithm for 2-approximation of matching.

(i) Let $M \leftarrow \emptyset$;

(ii) For any edge $e = (u, v)$ in the stream, add e to M if both of u, v are unmatched by M.

(iii) Return M.

We define a matching to be *maximal* (as opposed to maximum) if it is not a proper subset of any other matching. In other words, a maximal matching cannot be extended to a larger matching by just directly adding some other edges of the graph to it. It is easy to verify that, by construction, Algorithm 1 always outputs a maximal matching. We now prove that any maximal matching is a 2-approximation to the maximum matching which proves the correctness of the algorithm (the space of the algorithm is clearly $O(n)$ as it only stores a matching).

Lemma 4. *Any maximal matching M of a graph G has size $|M| \geq \mu(G)/2$.*

Proof. Let M^* be a maximum matching in G. For every edge (u, v) in M^* at least one of its endpoints is matched by M; otherwise, we could add (u, v) directly to M, violating its maximality. This means that the number of vertices matched by M is at least as large as the number of edges in M^*, i.e., $\mu(G)$. Since number of matched vertices is twice the edges, we get $|M| \geq \mu(G)/2$. $\square$

This algorithm is pretty standard and was first observed by Feigenbaum et.al. [7] who introduced the semi-streaming model for graph problems. Almost two decades later, we still do not know any better semi-streaming algorithms for this problem! On the lower bound front, Goel, Kapralov, and Khanna [9] proved that no semi-streaming algorithm can achieve a better than 3/2-approximation (we will see this result later in this lecture); this was subsequently improved by Kapralov to a better-than $(\frac{e}{e-1}) \sim 1.58$-approximation [13]. Closing the gap between the trivial 2-approximation upper bound and these lower bounds remains one of the most longstanding open problems in the graph streaming literature. To make progress on this fascinating question, several relaxations have been studied, such as:

- Studying the one-way communication complexity of bipartite matching; we will consider this in depth in the rest of this lecture.
- Considering random-arrival streams: Konrad, Maginez, and Mathieu [16] showed that one can beat the factor of 2 and achieved a 1.98-approximation. This setting was since studied extensively, culminating in a very recent semi-streaming algorithm of Bernstein [4] that achieves an almost (3/2)-approximation.
- Considering algorithms with one more pass: Here also, Konrad, Maginez, and Mathieu [16] showed that one can beat the factor of 2 and achieved a 1.92-approximation. These results has since been improved in a series of work, leading to an 1.71-approximation by [15]

3 One-Way Communication Complexity of Bipartite Matching

In order to understand single-pass semi-streaming algorithms for bipartite matching, Goel, Kapralov, and Khanna [9] initiated the study of bipartite matching in the one-way communication model. Recall that in this model, the edges of a bipartite graph $G = (L, R, E)$ are partitioned between Alice and Bob as E_A and E_B, respectively. Then, Alice sends a single message to Bob based on her input and Bob outputs an approximate maximum matching of G; what is the tradeoff between the size of messages and approximation ratio of the protocols? In particular, what is the best approximation ratio achievable by algorithms that communicate $\widetilde{O}(n)$ size messages. A lower bound in this model immediately imply a space lower bound for streaming algorithms and an upper bound gives insights and ideas for designing semi-streaming algorithms.

The main results of [9] are as follows: (i) there is a protocol that achieves a (3/2)-approximation to bipartite matching using $O(n)$ communication; and (ii) any better-than-(3/2)-approximation protocol requires $n^{1+\Omega(1/\log\log n)}$ communication[3]. This implies that there is no semi-streaming algorithm for (3/2)-approximation of bipartite matching and at the same time, proving stronger lower bounds requires other techniques that just one-way (two-player) communication complexity[4].

In the rest of this section, we go over the proof of each these results. For the upper bound part however, we present the algorithm of Assadi and Bernstein [2] instead that give a significantly simpler proof of this result (which unlike [9] also generalizes to general graphs) at a cost of increasing the approximation ratio to $(3/2 + \varepsilon)$ instead for any constant $\varepsilon > 0$.

3.1 A Communication Upper Bound for Bipartite Matching

We prove the following theorem.

Theorem 5 ([2])**.** *There is a one-way protocol for bipartite matching that for any $\varepsilon > 0$, achieves a $(3/2+\varepsilon)$-approximation using $O(n/\varepsilon)$ communication.*

The proof of this theorem is by showing that a particular *sparse* subgraph of Alice's input E_A, referred to as *edge-degree-constrained-subgraph (EDCS)*, can preserve "large" matchings approximately in a way that Bob can recover a large matching of $E_A \cup E_B$, given the EDCS and E_B instead. We start by presenting the required background on EDCS.

[3]The latter bound is larger than any $n \cdot \text{polylog}(n)$ as $n^{1/\log\log n} = 2^{\log n/\log\log n}$, while $\text{polylog}(n) = 2^{O(\log\log n)}$.

[4]The follow-up work of Kapralov [13] proves the $(\frac{e}{e-1})$-approximation lower bound by considering one-way communication complexity of matching when we have more than two-players; we will discuss multi-party communication models later in the course.

Edge-Degree-Constrained-Subgraphs (EDCS)

EDCS was defined first by Bernstein and Stein [5, 6] in the context of dynamic graph algorithms as follows:

Definition 6 ([5])**.** For any graph $G = (V, E)$ and parameters $\beta \geq 1$, $\varepsilon \in (0, 1)$, we define a (β, ε)-EDCS of G as any subgraph H satisfying the following two properties:

(i) For any $(u, v) \in H$, $\deg_H(u) + \deg_H(v) \leq \beta$;

(ii) For any $(u, v) \in G \setminus H$, $\deg_H(u) + \deg_H(v) \geq (1 - \varepsilon)\beta$.

Informally, edges in the EDCS should have a "low" edge-degree, while the edges missing from the EDCS should not have a "too low" edge degree.

Let us first prove that every graph admits an EDCS (with proper parameters)—this is non-trivial as the conditions of the EDCS are somewhat at odd with each other and thus a priori it is not clear whether one can satisfy both simultaneously[5].

Proposition 7 ([6])**.** *For any $\varepsilon > 0$ and $\beta \geq \frac{1}{\varepsilon}$, any graph $G = (V, E)$ (not necessarily bipartite) admits a (β, ε)-EDCS.*

Proof. Start with the empty graph H. While there exists an edge in H or $G \setminus H$ that violates properties (i) or (ii) of EDCS, respectively, fix this edge by removing it from H for the former or inserting it to H for the latter. We argue that this process terminates in a bounded number of steps. Define the potential function:

$$\Phi = \Phi(H) := (1 - \varepsilon/2) \cdot \beta \cdot \sum_{u \in V} \deg_H(u) - \sum_{(u,v) \in H} (\deg_H(u) + \deg_H(v)) \,.$$

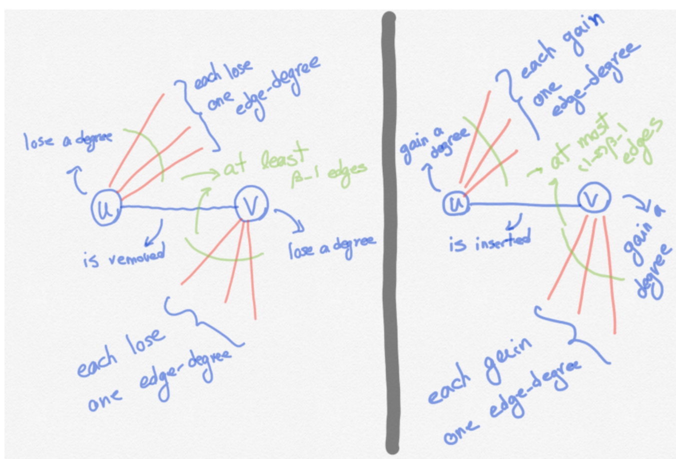

Figure 4: An illustration of the proof of existence of EDCS in Proposition 7.

After removing an edge (u, v) due to part (i), we have:

1. The first term decreases by $2\beta - \varepsilon\beta$ as two vertices lose a degree of one.
2. The second term (including the minus sign) increases by at least $\beta + 1$ (as an edge of degree $\geq \beta + 1$ is removed) plus another $\beta - 1$ (as each of the $\geq \beta - 1$ edges neighboring to (u, v) loses one degree).

Thus, in this case Φ increases by $\varepsilon\beta$. Similarly, after inserting an edge (u, v) due to part (ii), we have:

[5] In fact, not every graph admits a $(\beta, 0)$-EDCS (e.g., a star with two petals).

1. The first term increases by $2\beta - \varepsilon\beta$ as two vertices gained a degree of one.

2. The second term (including the minus sign) decreases by at most $(1-\varepsilon)\beta + 1$ (as an edge of prior degree $< (1-\varepsilon)\beta$ is inserted) plus another $(1-\varepsilon)\beta - 1$ (as each of the $\leq (1-\varepsilon)\beta - 1$ edges neighboring to (u, v) gain one degree).

Thus, in this case also Φ increases by $\varepsilon\beta$. Considering Φ starts from 0 and can only increase to $O(n\beta^2)$, this algorithm terminates in an EDCS after $O(n\beta/\varepsilon)$ fixing steps (the condition on β, ε ensures that $\varepsilon \cdot \beta \geq 1$). $\square$

We now prove the main property of EDCS in preserving matchings of the original graph approximately. This was first proved in [5] and was later simplified in [2]—we present the simpler proof here.

Proposition 8 ([5, 2])**.** *For any bipartite graph G and (β, ε)-EDCS H of G for $\varepsilon \in (0, 1/2]$ and $\beta \geq \frac{1}{\varepsilon}$,*

$$\mu(G) \leq (3/2 + 2\varepsilon) \cdot \mu(H).$$

Proof. Consider any Hall's theorem witness set A in H from Proposition 2 and assume $A \subseteq L$ by symmetry. We have $|A| - |N_H(A)| = n/2 - \mu(H)$. At the same time, in G, we have $|A| - |N_G(A)| \leq n/2 - \mu(G)$ (again by Proposition 2 otherwise maximum matching of G will be smaller than $\mu(G)$). Thus,

$$|N_G(A)| - |N_H(A)| \geq \mu(G) - \mu(H).$$

As $N_H(A) \subseteq N_G(A)$, the above bound means that there is a matching M in $G \setminus H$ of size $\mu(G) - \mu(H)$ between A and $R \setminus N_H(A)$ (the fact that this is a matching and not an arbitrary set of edges is because we can consider only the maximum matching of G in calculations above to the same effect). Let S be the endpoints of this matching. The following claim relates the size of S and $N_H(S)$.

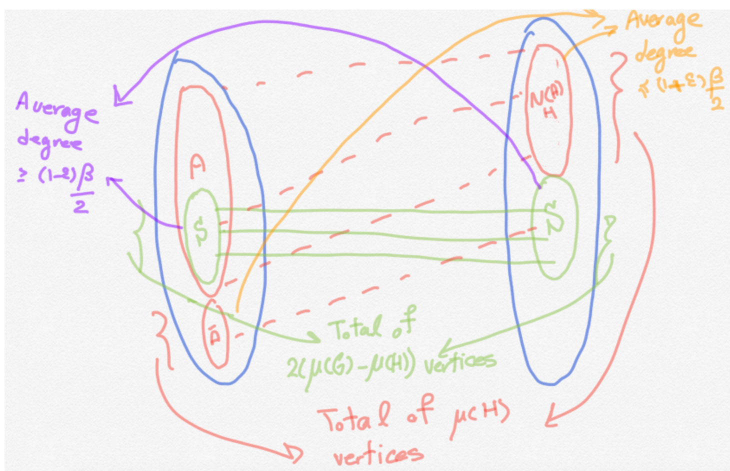

Figure 5: An illustration of the proof that EDCS preserves a large matching in Proposition 8.

Claim 9. $|S| \leq (1 + 4\varepsilon) \cdot |N_H(S)|$.

Proof. Let $\tilde{E} := E_H(S)$. We have,

$$\begin{aligned}
|\tilde{E}| &= \sum_{w \in S} \deg_H(w) && \text{(as no edge of } \tilde{E} = E_H(S) \text{ has both endpoints in } S) \\
&= \sum_{(u,v) \in M} \deg_H(u) + \deg_H(v) && \text{(by definition of } S \text{ being vertices of the matching } M) \\
&\geq (1-\varepsilon)\beta \cdot |M|. && \text{(by property } (ii) \text{ of EDCS for each edge of } M \subseteq G \setminus H) \\
&= (1-\varepsilon)\beta \cdot \frac{|S|}{2}. && (1)
\end{aligned}$$

On the other hand,

$$
\begin{aligned}
|\tilde{E}| \cdot \beta &\geq \sum_{(u,v)\in \tilde{E}} \deg_H(u) + \deg_H(v) \qquad \text{(by property (i) of EDCS for each edge of $\tilde{E} = E_H(S) \subseteq H$)} \\
&\geq \sum_{u \in S} \deg_{\tilde{E}}(u)^2 + \sum_{v \in N_H(S)} \deg_{\tilde{E}}(v)^2 \\
&\qquad \text{(as each vertex u contributes $\deg_{\tilde{E}}(u)$ terms to the sum each with a value of $\deg_{\tilde{E}}(u)$)} \\
&\geq |S| \cdot \left(\frac{|\tilde{E}|}{|S|}\right)^2 + |N_H(S)| \cdot \left(\frac{|\tilde{E}|}{|N_H(S)|}\right)^2 \\
&\text{($\textstyle\sum_{u\in S} \deg_{\tilde{E}}(u) = \sum_{v \in N_H(S)} \deg_{\tilde{E}}(u) = |\tilde{E}|$ and quadratic terms are minimized when all values are the same)} \\
&\geq |\tilde{E}| \cdot \frac{(1-\varepsilon)\beta}{2} + |\tilde{E}| \cdot \frac{|\tilde{E}|}{|N_H(S)|}. \qquad \text{(by Eq (1))}
\end{aligned}
$$

By reorganizing these terms, we obtain that,

$$
|N_H(S)| \geq \frac{|\tilde{E}|}{(1+\varepsilon)\beta/2} \geq \frac{(1-\varepsilon)\beta/2 \cdot |S|}{(1+\varepsilon)\beta/2} \geq \frac{1}{(1+4\varepsilon)} \cdot |S|,
$$

(by Eq (1) for second inequality and $\varepsilon \leq 1/2$ for the last)

as desired. □ Claim 9

We can now conclude the proof as follows. As A was a Hall's theorem witness in H, $N_H(S) \subseteq N_H(A) \cup L \setminus A$. Morever,

$$
|N_H(A) \cup L \setminus A| = |L| - (|A| - |N_H(A)|) = n/2 - (n/2 - \mu(H)) = \mu(H).
$$

Thus, $|N_H(S)| \leq \mu(H)$. On the other hand, $|S| = 2 \cdot (\mu(G) - \mu(H))$ and by Claim 9, we have,

$$
2(\mu(G) - \mu(H)) \leq (1+4\varepsilon) \cdot \mu(H),
$$

implying that $\mu(G) \leq (3/2 + 2\varepsilon) \cdot \mu(H)$ as desired. □

Remark. Let us briefly give an intuitive explanation of the proof of Proposition 8:

1. The Hall's theorem witness approach shows that there is a set S of $2(\mu(G) - \mu(H))$ vertices that are unmatched in H while there is a matching of this size between them in G. Property (ii) of EDCS implies that average degree of these vertices is at least $(1-\varepsilon)\beta/2$ in H.
2. Now this is where the "magic" of EDCS happens: for intuition, suppose all these vertices of S had degree exactly $(1-\varepsilon)\beta/2$ (instead of on average); then, the neighbors of these vertices in H, i.e., $N_H(S)$ should all have degree at most $(1+\varepsilon)\beta/2$ to satisfy property (i) of EDCS. This necessarily means that size of $N_H(S)$ can only be slightly smaller than S or informally $|S| \lesssim_\varepsilon |N_H(S)|$.
3. As all vertices in $N_H(S)$ are part of the matching $\mu(H)$, we get $2(\mu(G) - \mu(H)) \lesssim_\varepsilon \mu(H)$, which gives us the almost (3/2)-approximation.

The reader is strongly encouraged to consider applying this proof when H is instead a *maximal or even maximum* β-matching in G—a maximal/maximum subgraph of G with degree of every vertex bounded by β—and see why the same proof only gives a 2-approximation (which is also tight). *Hint:* Check what happens to the "magic" step above when working with vertex degrees instead of edge degrees.

Before moving on from this section, let us mention that Proposition 8 also holds for general graphs when $\beta = \Theta(1/\varepsilon^2 \log(1/\varepsilon))$; this was first proved in [6] and was simplified considerably in [2] (with slightly improved parameters).

The Protocol for Bipartite Matching

We are now ready to present our protocol in Theorem 5. The protocol basically involves Alice sending an EDCS of her input and Bob computing a maximum matching of this EDCS plus his own edges.

Algorithm 2. A one-way communication protocol for bipartite matching.

(i) Given E_A as input, Alice computes a $(\beta, \varepsilon/4)$-EDCS of E_A for the given parameter ε and $\beta = \lceil \frac{4}{\varepsilon} \rceil$ and sends it to Bob.

(ii) Given E_B as as input and H as the message of Alice, Bob outputs a maximum matching of $H \cup E_B$.

Algorithm 2 only communicates $O(n/\varepsilon)$ edges as EDCS H can only have $O(n\beta) = O(n/\varepsilon)$ edges by the property (i) of EDCS. The following lemma proves the correctness of this protocol.

Lemma 10. *Algorithm 2 outputs a $(3/2 + \varepsilon)$-approximation to the maximum matching of G.*

Proof. Let M^* be a maximum matching of G and M^*_B be the part of this matching that belongs to Bob's input (of course, this is unknown to the algorithm but this is only used for the analysis). Let $\tilde{M}$ be the following subset of M^*_B:

$$\tilde{M} := \{(u, v) \in M^*_B \mid \deg_H(u) + \deg_H(v) \le \beta\}.$$

We will prove that $H \cup \tilde{M}$ is a $(\beta + 2, \varepsilon/2)$-EDCS for $E_A \cup M^*_B$. This immediately concludes the proof as:

$$\mu(H \cup E_B) \ge \mu(H \cup \tilde{M}) \ge (3/2 + \varepsilon) \cdot \mu(E_A \cup M^*_B) = (3/2 + \varepsilon) \cdot \mu(G),$$

which hold because: (i) $H \cup E_B \supseteq H \cup \tilde{M}$, ($ii$) by Proposition 8, and (iii) $M^* \subseteq E_A \cup M^*_B$. It thus remains to prove that $\tilde{H} := H \cup \tilde{M}$ is indeed a $(\beta + 2, \varepsilon/2)$-EDCS for $E_A \cup M^*_B$. We do so by checking all EDCS properties (in the following, recall that H was a $(\beta, \varepsilon/4)$-EDCS of E_A):

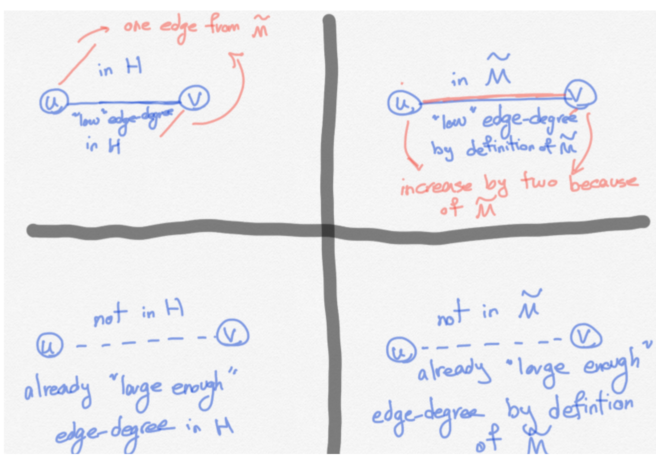

Figure 6: An illustration of the proof of Lemma 10.

- For any $(u, v) \in E_A$:
  - (i) If $(u, v) \in H$, then $\deg_{\tilde{H}}(u) + \deg_{\tilde{H}}(v) \le \beta + 2$ because degrees of vertices in H can only increase by one in $\tilde{H}$ and by property (i) of EDCS H;
  - (ii) If $(u, v) \notin H$, then $\deg_{\tilde{H}}(u) + \deg_{\tilde{H}}(v) \ge (1 - \varepsilon/4)\beta \ge (1 - \varepsilon/2)(\beta + 2)$ by property (ii) of EDCS.
- For any $(u, v) \in M^*_B$:

(*i*) If $(u, v) \in \tilde{M}$, then $\deg_{\tilde{H}}(u) + \deg_{\tilde{H}}(v) \leq \beta + 2$ because degrees of vertices in H can only increase by one in $\tilde{H}$ and by the definition of $\tilde{M}$;

(*ii*) If $(u, v) \notin \tilde{M}$, then $\deg_{\tilde{H}}(u) + \deg_{\tilde{H}}(v) > \beta > (1 - \varepsilon/2)(\beta + 2)$, again by the definition of $\tilde{M}$.

This means $\tilde{H}$ satisfies all properties of an $(\beta + 2, \varepsilon/2)$-EDCS for $E_A \cup M_B^*$, finalizing the proof. □

This concludes the proof of Theorem 5. We again remark that Theorem 5 even holds for general graphs (by increasing the communication to $O((n/\varepsilon^2) \cdot \log(1/\varepsilon))$; see [2].

3.2 A Communication Lower Bound for Bipartite Matching

Let us now switch to proving a tight lower bound on the approximation ratio achievable by protocols that use $\widetilde{O}(n)$ size messages. In particular, we prove the following theorem.

Theorem 11 ([9])**.** *For any constant $\varepsilon > 0$, any one-way protocol for bipartite matching with approximation ratio of $(3/2 - \varepsilon)$ requires $n^{1+\Omega_\varepsilon(1/\log\log n)}$ communication.*

The proof of this theorem is based on using a remarkable family of graphs that we define below.

Ruzsa-Szemerédi (RS) Graphs

A matching M in a graph $G = (V, E)$ is called an *induced matching* if there are no other edges between the vertices of this matching; in other words, the subgraph of G induced on vertices of M only consists of M.

Definition 12 (**Ruzsa-Szemerédi (RS) Graphs** [17])**.** We call a graph $G = (V, E)$ an (r, t)-RS graphs if its edges can be partitioned into t induced matchings $M_1, \ldots, M_t$, each of size r.

There is a trivial way of creating a bipartite (r, t)-RS graph with $t = (n/2r)^2$, and thus $r \cdot t = \Theta(n^2/r)$ edges:

(*i*) Let $L := L_1 \cup \cdots \cup L_{n/2r}$ and $R := R_1 \cup \cdots \cup R_{2n/r}$, with each component of size r.

(*ii*) Add a perfect matching M_{ij} of size r between any two L_i and R_j.

This graph forms an (r, t)-RS graph as there are no other edges between $L_i \times R_j$ beside the matching M_{ij} of size r and $t = (n/2r)^2$. For our application, we are interested in the case when $r = \Theta(n)$; this construction, at this point, only gives a very sparse (r, t)-RS graph with $\Theta(n)$ edges. Let us first see that we can create a bipartite (r, t)-RS graph with $r = n/4$ (so half the size of a perfect matching) and $\Theta(n \log n)$ edges.

A slightly non-trivial RS graph. Consider the following graph:

(*i*) Let $L := \{0,1\}^k$ and $R := \{0,1\}^k$ for $k = \log(n/2) = \log n - 1$ (so $|L| = |R| = n/2$). We identify the vertices on each side of the bipartition by a k-bit string $x_1 x_2 \ldots x_k$.

(*ii*) For any $i \in [k]$, add a matching M_i defined as follows to the graph:

$$M_i := \left\{ (u, v) \mid u = x_1 x_2 \cdots \underbrace{0}_{i} \cdots x_{k-1} x_k \in L \text{ and } v = x_1 x_2 \cdots \underbrace{1}_{i} \cdots x_{k-1} x_k \in R \right\}.$$

This graph has n vertices and $(\log n - 1)$ matchings of size $n/4$ each (as each M_i matches half of L to a half of R). To see that these matchings are induced, consider two edges $(u_1, v_1), (u_2, v_2) \in M_i$; we claim that there is no other matching M_j such that $(u_1, v_2) \in M_j$. Suppose this is not the case by way of contradiction. Since $u_1 \in L$ is matched by both M_i and M_j, we have that both i-th index and j-th index of its string is 0. On the other hand, since $v_2 \in R$ is also matched by both M_i and M_j, we have that both its i-th index and j-th index are 1. By then M_j cannot have the edge (u_1, v_2) as any edge in the matchings of this graph only change one index of the endpoint vertices' strings.

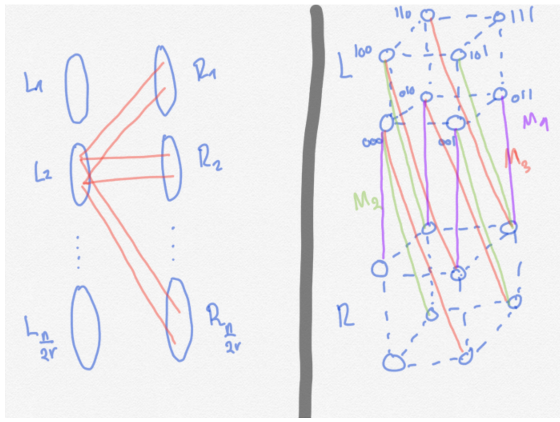

Figure 7: An illustration of the trivial (left) and the slightly non-trivial (right) RS graphs.

A much more sophisticated RS graph. It turns out that one can do much better than this construction in terms of the density of the RS graphs. In particular, Fischer et.al. [8] designed an (r, t)-RS graph with $r = n/6$ and $t = n^{\Omega(1/\log\log n)}$; this was strengthened in [9] to parameters $r = (1/4 - \varepsilon)n$ and $t = n^{\Omega_\varepsilon(1/\log\log n)}$, leading to the following proposition.

Proposition 13 ([9, 8])**.** *For any constant $\varepsilon > 0$, there is a bipartite (r, t)-RS graph with $r = (1/4 - \varepsilon)n$ and $t = n^{\Omega_\varepsilon(1/\log\log n)}$, hence $n^{1+\Omega(1/\log\log n)}$ edges.*

We will not go over the details of this construction in this lecture and instead simply use it directly to prove the communication lower bound.

Remark. (Dense) RS graphs are in general highly fascinating objects: they are *locally sparse* as they are formed via induced matchings[a], while being *globally dense*. For some range of parameters, there are surprising construction of these graphs: for instance, Alon, Moitra, and Sudakov [1] proved that there are (r, t)-RS graphs with $r = n^{1-o(1)}$ and $\binom{n}{2} - o(n^2)$ edges! (compare this with the bounds obtained by the trivial RS graph). We refer the interested reader to [1] for more background on RS graphs.

[a]For any induced matching of size r, there are roughly r^2 edges that can no longer belong to the graph.

The Lower Bound for Matching based on RS Graphs

We now prove Theorem 11. The proof uses Yao's minimax principle by analyzing deterministic algorithms over a fixed distribution of inputs. The input distribution is informally as follows: Alice is given an RS graph and Bob is given a matching; they are chosen in a way that only one induced matchings of Alice is "important" for getting a better than (3/2)-approximation (using the fact that the matchings are induced); however, Alice is oblivious to this matching and thus needs to communicate about all her edges in order to communicate about the special one as well. We now formalize this as follows.

A hard distribution μ. Let $G^{\mathsf{rs}} := (L^{\mathsf{rs}}, R^{\mathsf{rs}}, E^{\mathsf{rs}})$ an (r, t)-RS graph with N vertices, $r = (N/4 - \varepsilon N)$ and $t = N^{\Omega_\varepsilon(1/\log\log N)}$, and induced matchings $M_1^{\mathsf{rs}}, \ldots, M_t^{\mathsf{rs}}$ (guaranteed to exists by Proposition 13). This RS graph is *known* to both players.

- (i) From every induced matching M_i^{rs}, we remove $\varepsilon \cdot N$ edges independently and uniformly at random and given the remaining edges as M_i to Alice. (So, $E_A := M_1 \cup \cdots \cup M_t$).
- (ii) We pick one induced matching M_j^{rs} independently and uniformly at random (called the *special* matching). Bob gets a perfect matching between two new sets of vertices L_B and R_B, each of size $(N/4+\varepsilon N)$ and the vertices of G^{rs} that are *not* in M_j^{rs}.

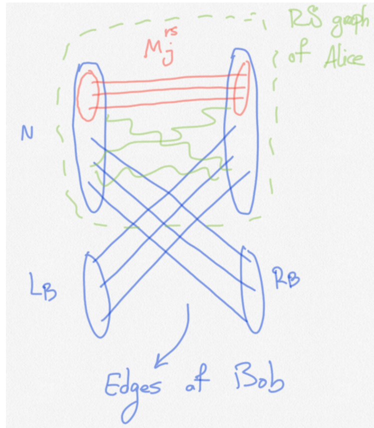

Figure 8: An illustration of the graphs in the hard distribution μ.

We note that the graphs constructed in μ have $n = N + N/2 + 2\varepsilon N$ vertices (N vertices from the RS graph and $N/2 + 2\varepsilon N$ from the extra vertices that only Bob has an edge to). The following lemma identifies the "main task" of protocols when for approximating bipartite matching on graphs sampled from μ.

Lemma 14. *In every graph G sampled from μ:*

(i) *There is a matching of size $\mu(G) \geq 3N/4$;*

(ii) *Any matching M of size $N/2 + 2\varepsilon N + b$ edges in G has b edges from the special matching M_j.*

Proof. We prove each part separately as follows:

(i) We can pick $N/2 + 2\varepsilon N$ edges of Bob plus $N/4 - 2\varepsilon N$ edge of Alice in $M_j \subset M_j^{\mathsf{rs}}$ in every graph of μ to form a matching of the desired size.

(ii) Without loss of generality, we can assume M consists of all edges of Bob of size $N/2 + 2\varepsilon N$ edges (as vertices in L_A and L_B only have degree one) and thus the remaining unmatched vertices are only between endpoints of M_j^{rs}; since this is an induced matching, the extra b edges picked by M should belong to M_j.

The lemma now follows from the two parts above. $\square$

Now consider a deterministic protocol π for matching on graphs of μ and suppose its communication cost is $c = o(r \cdot t)$. We assume that π never outputs an edge in the matching that does not belong to the input graph (thus it may err by only outputting a not-large-enough matching but not a "wrong" one)[6]. Lemma 14 suggests that the task of π is simply to convey which edges of M_j^{rs} actually belong to the matching M_j of Alice so that Bob can output them. However, as Alice is oblivious to the identity of j, she effectively needs to communicate such information about all matchings $M_1^{\mathsf{rs}}, \ldots, M_t^{\mathsf{rs}}$. We now formalize this as follows.

Definition 15. For a message Π sent by the protocol π, we use $\mathcal{G}(\Pi)$ to denote the set of all graphs of Alice that are mapped to the same message Π. We say that an edge e *belongs to* $\mathcal{G}(\Pi)$ if $e \in G$ for all $G \in \mathcal{G}(\Pi)$. For every $i \in [t]$, we use $M(\Pi)_i$ to denote the set of all edges in M_i^{rs} that belong to Π.

The discussion earlier about the correctness of π implies that Bob, given message Π, can only output an edge e in the final matching if e belongs to $\mathcal{G}(\Pi)$. Combined with Lemma 14, this implies that ratio of the

[6]This is a natural assumption made in [9] but it is not without loss of generality. However, one can easily lift this assumption using a slightly more careful analysis (or alternatively a reduction from Index) and thus we ignore this issue for now.

size of the matching output by Bob to maximum matching of G is at most

$$\frac{N/2 + 2\varepsilon N + |M(\Pi)_j|}{3N/4} \leq 2/3 + 3\varepsilon + \frac{|M(\Pi)_j|}{N}. \tag{2}$$

The following lemma is the main part of the proof that bounds the last term in the RHS above.

Lemma 16. *W.p.* 0.8, $|M(\Pi)_j| \leq \varepsilon N$.

Proof. We start by claiming that with high probability, $\mathcal{G}(\Pi)$ is going to be "large".

Claim 17. *With probability* $1 - o(1)$ *over* Π, $|\mathcal{G}(\Pi)| \geq \binom{r}{\varepsilon N}^t / 2^{2c}$, *where c is the communication cost of π.*

Proof. The total number of input edges to Alice is $\binom{r}{\varepsilon N}^t$ and these are partitioned into 2^c different sets $\mathcal{G}(\Pi)$ for all the 2^c choices of Π. Moreover, the probability that a message Π is sent by Alice is exactly $|\mathcal{G}(\Pi)| / \binom{r}{\varepsilon N}^t$ as the distribution of inputs are uniform. As such, by a union bound over the 2^c choices of Π,

$$\Pr_{\Pi}\left(|\mathcal{G}(\Pi)| < \frac{\binom{r}{\varepsilon N}^t}{2^{2c}}\right) \leq 2^c \cdot \frac{\binom{r}{\varepsilon N}^t}{2^{2c} \cdot \binom{r}{\varepsilon N}^t} = o(1),$$

as desired. $\square$ Claim 17

We refer to any Π that satisfies the event of Claim 17 as a *large* message. We now prove that for any large message Π, $M(\Pi)_i$ for "most" indices i is "very small".

Claim 18. *For any large message* Π, *the number of indices* $i \in [t]$ *such that* $M(\Pi)_i < \varepsilon N$ *is at least* $0.9 \cdot t$.

Proof. Suppose towards a contradiction that for more that $0.1 \cdot t$ indices $i \in [t]$, we have $M(\Pi)_i \geq \varepsilon N$. Then,

$$\begin{aligned} |\mathcal{G}(\Pi)| &\leq \binom{r}{\varepsilon N}^{0.9t} \cdot \binom{r - \varepsilon N}{\varepsilon N}^{0.1t} \\ &\qquad \text{(because edges of } M(\Pi)_i \text{ cannot be part of the } \varepsilon N \text{ edges removed from } M_i^{\mathsf{rs}} \text{ in } M_i) \\ &\leq \binom{r}{\varepsilon N}^{t} \cdot 2^{-\Omega(N \cdot t)} \qquad \text{(because } \tbinom{a-b}{c} \leq 2^{-\Theta(b)}\tbinom{a}{c} \text{ for } c = \Theta(a)) \end{aligned}$$

But this is smaller than the lower bound on size of $|\mathcal{G}(\Pi)|$ guaranteed by Claim 17 whenever $c = o(r \cdot t) = o(N \cdot t)$, a contradiction.

We are now done because the choice of the special index $j \in [t]$ as input to Bob is independent of the choice of the graph given to Alice and thus her message Π. As such, by Claim 17, w.p. $1 - o(1)$, we will have a large Π and by Claim 18, w.p. 0.9 over the choice of j, $M(\Pi)_i < \varepsilon N$; a union bound, finalizes the proof. $\square$

By Lemma 16 and Eq (2), w.p. at least $4/5 > 2/3$, the approximation ratio of the protocol is at least $(2/3 + 4\varepsilon)^{-1} \geq (3/2 - 12\varepsilon)$. By re-parameterizing ε with $\varepsilon' = \varepsilon/12$ in the arguments above, we obtain that any $(3/2 - \varepsilon')$-approximation requires $n^{1+\Omega(1/\log\log n)}$, finalizing the proof of Theorem 11.

References

[1] N. Alon, A. Moitra, and B. Sudakov. Nearly complete graphs decomposable into large induced matchings and their applications. In *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pages 1079–1090, 2012. 11

[2] S. Assadi and A. Bernstein. Towards a unified theory of sparsification for matching problems. In *2nd Symposium on Simplicity in Algorithms, SOSA@SODA 2019, January 8-9, 2019 - San Diego, CA, USA*, pages 11:1–11:20, 2019. 5, 7, 8, 10

[3] S. Assadi and R. Raz. Near-quadratic lower bounds for two-pass graph streaming algorithms. *CoRR*, abs/2009.01161. To appear in FOCS 2020, 2020. 4

[4] A. Bernstein. Improved bounds for matching in random-order streams. In *47th International Colloquium on Automata, Languages, and Programming, ICALP 2020, July 8-11, 2020, Saarbrücken, Germany (Virtual Conference)*, pages 12:1–12:13, 2020. 5

[5] A. Bernstein and C. Stein. Fully dynamic matching in bipartite graphs. In *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part I*, pages 167–179, 2015. 6, 7

[6] A. Bernstein and C. Stein. Faster fully dynamic matchings with small approximation ratios. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 692–711, 2016. 6, 8

[7] J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, and J. Zhang. On graph problems in a semi-streaming model. In *Automata, Languages and Programming: 31st International Colloquium, ICALP 2004, Turku, Finland, July 12-16, 2004. Proceedings*, pages 531–543, 2004. 3, 4

[8] E. Fischer, E. Lehman, I. Newman, S. Raskhodnikova, R. Rubinfeld, and A. Samorodnitsky. Monotonicity testing over general poset domains. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 474–483, 2002. 11

[9] A. Goel, M. Kapralov, and S. Khanna. On the communication and streaming complexity of maximum bipartite matching. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012*, pages 468–485, 2012. 4, 5, 10, 11, 12

[10] V. Guruswami and K. Onak. Superlinear lower bounds for multipass graph processing. In *Proceedings of the 28th Conference on Computational Complexity, CCC 2013, K.lo Alto, California, USA, 5-7 June, 2013*, pages 287–298, 2013. 4

[11] P. Hall. On representatives of subsets. In *Classic Papers in Combinatorics*, pages 58–62. Springer, 2009. 2

[12] J. E. Hopcroft and R. M. Karp. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM J. Comput.*, 2(4):225–231, 1973. 1

[13] M. Kapralov. Better bounds for matchings in the streaming model. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 1679–1697, 2013. 5

[14] D. König. Über graphen und ihre anwendung auf determinantentheorie und mengenlehre. *Mathematische Annalen*, 77(4):453–465, 1916. 1

[15] C. Konrad. A simple augmentation method for matchings with applications to streaming algorithms. In *43rd International Symposium on Mathematical Foundations of Computer Science, MFCS 2018, August 27-31, 2018, Liverpool, UK*, pages 74:1–74:16, 2018. 5

[16] C. Konrad, F. Magniez, and C. Mathieu. Maximum matching in semi-streaming with few passes. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 15th International Workshop, APPROX 2012, and 16th International Workshop, RANDOM 2012, Cambridge, MA, USA, August 15-17, 2012. Proceedings*, pages 231–242, 2012. 5

[17] I. Z. Ruzsa and E. Szemerédi. Triple systems with no six points carrying three triangles. *Combinatorics (Keszthely, 1976), Coll. Math. Soc. J. Bolyai*, 18:939–945, 1978. 10

[18] J. van den Brand, Y.-T. Lee, D. Nanongkai, R. Peng, T. Saranurak, A. Sidford, Z. Song, and D. Wang. Bipartite matching in nearly-linear time on moderately dense graphs. *CoRR*, abs/2009.01802. To appear in FOCS 2020, 2020. 1