Lower Bounds for Distributed Sketching of Maximal Matchings and Maximal Independent Sets

Sepehr Assadi
sepehr.assadi@rutgers.edu
Rutgers University

Gillat Kol
gillat.kol@gmail.com
Princeton University

Rotem Oshman
roshman@tau.ac.il
Tel Aviv University

Abstract

Consider the following distributed graph sketching model: There is a referee and n vertices in an undirected graph G sharing public randomness. Each vertex v only knows its neighborhood in G and the referee receives no input initially. The vertices simultaneously each sends a message, called a *sketch*, to the referee who then based on the received sketches outputs a solution to some combinatorial problem on G, say, the minimum spanning tree problem.

Previous work on graph sketching have shown that numerous problems, including connectivity, minimum spanning tree, edge or vertex connectivity, cut or spectral sparsifiers, and $(\Delta+1)$-vertex coloring, all admit efficient algorithms in this model that only require sketches of size polylog(n) per vertex. In contrast, we prove that the two fundamental problems of maximal matching and maximal independent set do *not* admit such efficient solutions: Any algorithm for either problem that errs with a small constant probability requires sketches of size $\Omega(n^{1/2-\varepsilon})$ for any constant $\varepsilon > 0$.

We prove our results by analyzing communication complexity of these problems in a communication model that allows sharing of inputs between limited number of players, and hence lies between the standard number-in-hand and number-on-forehead multi-party communication models. Our proofs are based on a family of hard instances using Ruzsa-Szemerédi graphs and information-theoretic arguments to establish the communication lower bounds.

CCS Concepts

• **Theory of computation → Distributed algorithms**.

Keywords

Distributed sketching, maximal matching, maximal independent set, communication complexity, broadcast congested clique

ACM Reference Format:

Sepehr Assadi, Gillat Kol, and Rotem Oshman. 2020. Lower Bounds for Distributed Sketching of Maximal Matchings and Maximal Independent Sets. In *ACM Symposium on Principles of Distributed Computing (PODC '20), August 3–7, 2020, Virtual Event, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3382734.3405732

PODC '20, August 3–7, 2020, Virtual Event, Italy

ACM ISBN 978-1-4503-7582-5/20/08…$15.00
https://doi.org/10.1145/3382734.3405732

1 Introduction

We consider the following distributed sketching model: There are n vertices indexed by $[n]$ in an undirected graph $G(V,E)$ and we want to solve some combinatorial problem P on G, say, find a spanning forest of G. Any given vertex v only knows its own index and the set of indices of its neighbors denoted by $N(v)$. The vertices also have access to a shared random string referred to as public coins. Then, each vertex v sends a message—called a *sketch* sk(v)—to a referee, who based on the received sketches and the public coins must output a solution to $P(G)$ with constant probability. The task is to minimize the size of the sketches measured in number of bits (the problem is trivial with sketches of size $\Theta(n)$ by sending the entire neighborhood of each vertex to the referee).

At first glance, it may not be clear that this model allows for interesting solutions to non-trivial graph problems. For instance, consider the spanning forest problem and suppose the input graph consists of two disjoint random graphs connected by an edge (u,v). Clearly edge (u,v) is part of any spanning forest but from the perspective of vertices u and v, edge (u,v) is indistinguishable from their other edges. This seems to suggest that unless sk(u) or sk(v) is of size $\Omega(n)$, the referee should not be able to find (u,v). This intuition is however *not* correct: since each edge in this model is seen by both its endpoints, vertices other than u and v can also "inform" the referee about other edges incident on u and v. Hence, by combining this information with sketches of u and v, we should be able to use much smaller sketches and still allow the referee to recover the edge (u,v)[1]. Indeed, an elegant algorithm by [1], referred to as AGM sketches, shows that for finding spanning forest of any given graph with high probability, we only need messages size $O(\log^3 n)$ bits.

Starting from the AGM sketches of [1], there has been tremendous progress in obtaining efficient graph sketching algorithms for various problems, including minimum spanning trees and edge connectivity [1], subgraph counting [2], vertex connectivity [37], cut sparsifiers and approximate min/max cuts [2], spectral sparsifiers [3, 43], densest subgraph [22, 48], graph degeneracy [31], and $(\Delta + 1)$ vertex coloring [11]. Despite this however, obtaining similarly efficient sketches for the two fundamental and closely related problems of maximal matching and maximal independent set has remained elusive. Our goal in this work is to address this gap in our understanding of these two key problems.

[1]For the interested reader, here is a concrete solution to this particular example. Firstly, sending $O(\log n)$ incident edges uniformly at random per vertex ensures that the referee can identify the partition of vertices w.h.p. Each vertex w also computes the number $s_w := \sum_{z\in N(w):z>w}(z\cdot n+w) - \sum_{z\in N(w):z<w}(w\cdot n+z)$ and sends it to the referee. The referee then sums up all the numbers sent by vertices inside one of the partitions: it is easy to see that the value of this sum uniquely identifies the edge (u,v) as the contribution of all edges inside the partition cancels out.

1.1 Our Contributions

We prove that in contrast to all the aforementioned problems, neither maximal matching nor maximal independent set admit efficient polylog(n) size sketches in the distributed sketching model.

Result 1. *Any public-coin distributed sketching protocol for computing a maximal matching or a maximal independent set with constant probability of success requires* $\Omega(n^{1/2-\varepsilon})$ *size sketches for any constant* $\varepsilon > 0$.

Result 1 should be contrasted in particular with the complexity of another closely related symmetry breaking problem, the $(\Delta + 1)$ coloring problem, that admits sketches of size $O(\log^3 n)$ bits [11]. Result 1 can also be interpreted directly as a lower bound in the *broadcast congested clique* model for one-round algorithms (see, e.g. [30, 39] for the definition of this model and in particular its equivalence to the distributed sketching model).

It is also worth comparing Result 1 to the lower bounds for these two problems in the related *streaming* model. Assadi *et al.* [14] have shown a lower bound for approximate matching in dynamic graph streams and Assadi *et al.* [11] and Cormode *et al.* [26] have proven a lower bound for maximal independent set in insertion-only streams. By straightforward reductions (see, e.g. [1]), these results imply that any distributed sketching algorithm that uses *linear* sketches require $\Omega(n)$ size sketches for either problem (a linear sketch is a *linear* transformation of the input of players as opposed to an arbitrary sketch). However, unlike our Result 1, these results do *not* imply any lower bounds for general sketches (see, e.g. [8], for a separation between general vs linear sketches).

Finally, Result 1 leaves a gap of roughly $n^{1/2}$ between the lower bound and the trivial upper bound of $O(n)$. Closing this gap remains an interesting open question. We note that even though we are not aware of any better upper bound for either problem in our model, *if* one allows only *one extra* round of sketching, then both problems admit (adaptive) sketches of size $O(n^{1/2})$ by results of [46] and [35] for maximal matching and maximal independent set, respectively.

1.2 Our Techniques

The starting point of our work is the lower bound approach of [14] for *linear* sketches of approximate matching. In [14], the authors gave a communication complexity lower bound for approximating matching in the following communication model: The input graph is *edge-partitioned* between a small number of $n^{o(1)}$ players and the players need to simultaneously send a message to the referee to solve the problem. By picking the input graph of each player to *locally* be a dense Ruzsa-Szemerédi graph (a graph with a "large" number of "large" *induced* matchings; see Section 2.2) that are "incompressible" in the context of matching problem, [14] manages to ensure that the players need to communicate almost their entire graph to the referee in order to compute a large matching.

To lift this approach to our model, we need to address several key aspects of the distributed sketching model that are missing from the communication model of [14]. Firstly, the input graph in our model is *vertex-partitioned* between the players in that each player gets to see *all* edges incident on a vertex. This property right away breaks the "incompressibility"-type arguments in [14] based on Ruzsa-Szemerédi graphs as seeing *all* edges incident on vertices allows *some* of the players to figure out on their own which induced matching in the Ruzsa-Szemerédi graph is the *important* one and solely focus on communicating edges of that matching. Secondly, since each edge is seen by *both* its endpoint in our model, i.e., the inputs are shared in a limited way, a player can inform the referee about the edges of another player as well (a simple example is outlined in Footnote 1). Combining this with the first challenge above means that we will have some players that not only know which parts of the graph are more important to focus on and communicate to the referee, but can also inform the referee about the input of *other* players in those parts of the graph!

We manage to address the challenges above through a combination of ideas. We first change the input distribution of [14] in order to limit the number of players that have extra knowledge about the important parts of the graph (which we call *public* players). The main step is then to "decompose" the information revealed by messages of players to the referee between the public players and non-public players and bound each part separately. This requires entirely foregoing the combinatorial arguments in lower bound of [14] and instead use information-theoretic tools for the analysis of the lower bound. Imposing the limit on the number of public players then allow us to argue that even though they have a good knowledge of which parts of the graph to communicate, their total bandwidth is not enough for solving the problem on their own. Finally, we show that the non-public players will not be able to communicate much about their important edges with low communication as they are unaware of the identity of their important edges and combine these to finalize the proof.

1.3 Related Work

To our knowledge, the distributed sketching model we study in this paper was first considered by Becker *et al.* in [17, 18]. In particular, [17] proved lower bounds for deterministic algorithms for computing some local properties of the graph such as triangle-freeness and [18] extended some of these lower bounds to randomized algorithms. Moreover, [18] proved separations between power of deterministic, private-coin, and public-coin algorithms. Designing algorithms in this and related graph sketching models has been a subject of extensive study after the breakthrough result of Ahn, Guha, and McGregor [1] on obtaining an $O(\log^3 n)$ size sketches for the spanning tree problem which paved the path for various algorithmic results mentioned earlier. Finally, on the lower bound front, Nelson and Yu [50], building on [44], proved that any public-coin problem for the spanning tree problem requires $\Omega(\log^3 n)$ size sketches. Proving super-logarithmic lower bounds for the spanning tree problem for private-coin or deterministic protocols remains a fascinating open problem in this area [21].

The distributed sketching model in our paper is equivalent to the broadcast congested clique model when restricted to one-round protocols. This model has been studied in several recent papers from both upper and lower bounds perspectives; see, e.g. [16, 30, 39–41] and references therein. For instance, Jurdzinski and Nowicki study deterministic algorithms for graph connectivity in this model [40, 41], Becker *et al.* [19] consider algorithms and lower bounds for finding small cycles, Montealegre *et al.* [49] study reconstruction of hereditary graph classes, and Drucker *et al.* [30] prove lower bounds for multi-round algorithms for several problems including testing triangle-freeness or K_4-freeness.

Another model similar to our setting is when the input graph is bipartite and there is a player for each vertex in *only one side* of the bipartition [6, 9, 24, 29] (this model has application to algorithmic game theory where players correspond to bidders in an auction and the other side of the graph are items they are interested in). Unlike our model, this setting no longer has *shared* inputs between the players and strong lower bounds are known in this problem for problems such as approximating matching [6, 24, 29] and even computing a spanning forest which is an "easy" problem in our model. Roughly speaking, the source of hardness in all these lower bounds are vertices of degree *one* on the non-player-side of the bipartition that are hard to find for the player-side. When allowing all vertices to send a message in our model, these degree one vertices can easily identify themselves and break the lower bound.

We refer the interested reader to [1, 2, 17, 30, 33, 39, 50] for further discussion of related work and connection of this model to related models such as CONGEST and dynamic streams.

2 Notation and Preliminaries

For any integer $t \in \mathbb{N}$, we use $[t] := \{1, \ldots, t\}$. For a graph $G = (V, E)$, $n = |V|$ denote the number of vertices and $m = |E|$ denote the number of edges. We use san-serif fonts to denote random variables to avoid ambiguity with the value they can take.

2.1 Communication Model

The communication model we work with can be defined formally as follows. Consider an undirected graph $G = (V, E)$. There is one player per every vertex of the graph and a central referee (or coordinator). The input to player corresponding to vertex $u \in V$ is the number of vertices n, the ID of node u which is a unique integer in $\{1, \ldots, n\}$, and the set of all neighbors v of u in G or alternatively all edges $(u, v) \in E$. It is important to emphasize that any edge $(u, v) \in E$ is thus given as input to *two* players, namely, u and v. The referee receives no input.

In this paper, we are interested in computing a maximal matching or a maximal independent set (MIS) of the graph G. In order to do this, each player is allowed to, simultaneously with other players, send a single message to the referee based solely on the player's input, who upon receiving the input messages computes the final output. A *protocol* in this model describes the algorithms of players (for computing the messages) and the algorithm of the referee (for recovering the solution from received messages). We define *communication cost* of a protocol as the *worst case* length of the message sent by any player in the protocol (measured in number of bits). For randomized protocols, we allow the players and the referee to have access to *public-coins*, i.e., a *shared* random string that can be used by the algorithms of players and the referee.

Types of error: Naturally, we allow randomized protocols to make error (with some fixed probability). This means a protocol for maximal matching may err by outputting a matching that contains an edge not in the graph, or a matching which is not maximal. Similarly, a protocol for maximal independent set may err by outputting a set which is not an independent set or is not maximal.

We shall note that many lower bounds in the literature for approximate matching, e.g. in [12, 36, 42, 45] make this implicit assumption that the output of the protocol is *always* a valid matching (but may not necessarily be sufficiently large) which weakens the lower bound. Moreover, in order for our reduction for maximal independent set to work, we truly need to prove the lower bound for matching algorithms that are allowed outputting edges that may not be part of the graph with some small error probability.

We point out that the communication model studied in this paper lies between the two key multiparty communication models, the *number-in-hand (NIH) model* (in which the inputs of players are disjoint) and the *number-on-forehead (NOF) model* (in which the inputs of players can be arbitrarily overlapping). Compared to the NIH model, proving communication complexity lower bounds in the NOF model are considerably more challenging (see, e.g. [15, 25, 47]).

2.2 Ruzsa-Szemerédi Graphs

A graph $G^{\mathsf{RS}}(V, E)$ is a called an (r, t)-Ruzsa-Szemerédi graph (RS graph for short) iff its edge-set E can be partitioned into t *induced matchings* $M_1^{\mathsf{RS}}, \ldots, M_t^{\mathsf{RS}}$, each of size r. We use the original construction of RS graphs due to Ruzsa and Szemerédi [51], based on the existence of large sets of integers with no 3-term arithmetic progression, proven by Behrend [20] (we note that there are multiple other constructions with different parameters; see, e.g. [5, 32, 34, 36] and references therein).

Proposition 2.1 ([51]). *For infinitely many integer N, there are (r, t)-RS graphs on N vertices with $r = \frac{N}{e^{\Theta(\sqrt{\log N})}}$ and $t = N/3$.*

RS graphs have been extensively studied as they arise naturally in property testing, PCP constructions, additive combinatorics, streaming algorithms, graph sparsification, etc. (see, e.g., [4, 5, 7, 10, 13, 14, 23, 26, 28, 32, 34, 36, 38, 42, 45, 47, 52]). In particular, a line of work initiated by Goel, Kapralov, and Khanna [36] have used different constructions of these graphs to prove communication complexity lower bounds for (approximate) matching algorithms in different settings [13, 14, 26, 36, 42, 45].

2.3 Basic Information Theory Facts

Our proof relies on basic concepts from information theory which we summarize below. We refer the interested reader to the excellent text by Cover and Thomas [27] for a broader introduction.

For random variables A, B, we use $\mathbb{H}(\mathsf{A})$ and $\mathbb{I}(\mathsf{A}\,;\mathsf{B})$ to denote the Shannon entropy and mutual information, respectively. We shall use the following basic properties of entropy and mutual information in the paper.

Fact 2.2. *Let A, B, C, and D be four random variables.*

(1) $0 \leq \mathbb{H}(\mathsf{A}) \leq \log|\mathrm{supp}(\mathsf{A})|$ *(where* $\mathrm{supp}(\mathsf{A})$ *denote the support of* A*). The right equality holds iff* A *is uniformly distributed.*
(2) $\mathbb{I}(\mathsf{A}\,;\mathsf{B}) \geq 0$. *The equality holds iff* A *and* B *are* independent.
(3) Conditioning on a random variable reduces entropy*:* $\mathbb{H}(\mathsf{A} \mid \mathsf{B}, \mathsf{C}) \leq \mathbb{H}(\mathsf{A} \mid \mathsf{B})$. *The equality holds iff* $\mathsf{A} \perp \mathsf{C} \mid \mathsf{B}$.

(4) Chain rule for entropy*:* $\mathbb{H}(A, B \mid C) = \mathbb{H}(A \mid C) + \mathbb{H}(B \mid C, A)$.
(5) Chain rule for mutual information*:* $\mathbb{I}(A, B\,; C \mid D) = \mathbb{I}(A\,; C \mid D) + \mathbb{I}(B\,; C \mid A, D)$.

We will use the following two standard inequalities regarding the effect of conditioning on mutual information.

Proposition 2.3. *If* $A \perp D \mid C$*:* $\mathbb{I}(A\,; B \mid C) \leq \mathbb{I}(A\,; B \mid C, D)$.

Proof. By Fact 2.2-(3), since $A \perp D \mid C$, we have $\mathbb{H}(A \mid C) = \mathbb{H}(A \mid C, D)$ and since conditioning can only decrease the entropy, $\mathbb{H}(A \mid C, B) \geq \mathbb{H}(A \mid C, B, D)$. As such,

$$\mathbb{I}(A\,; B \mid C) = \mathbb{H}(A \mid C) - \mathbb{H}(A \mid C, B) \\ \leq \mathbb{H}(A \mid C, D) - \mathbb{H}(A \mid C, B, D) = \mathbb{I}(A\,; B \mid C, D),$$

concluding the proof. ∎

Proposition 2.4. *If* $A \perp D \mid B, C$*:* $\mathbb{I}(A\,; B \mid C) \geq \mathbb{I}(A\,; B \mid C, D)$.

Proof. By Fact 2.2-(3), since $A \perp D \mid B, C$, we have $\mathbb{H}(A \mid B, C) = \mathbb{H}(A \mid B, C, D)$ and since conditioning can only reduce the entropy, $\mathbb{H}(A \mid C) \geq \mathbb{H}(A \mid D, C)$. As such,

$$\mathbb{I}(A\,; B \mid C) = \mathbb{H}(A \mid C) - \mathbb{H}(A \mid B, C) \\ \geq \mathbb{H}(A \mid D, C) - \mathbb{H}(A \mid B, C, D) = \mathbb{I}(A\,; B \mid C, D),$$

concluding the proof. ∎

3 A Lower Bound for Maximal Matching

We prove the following theorem in this section, which implies Result 1 for matching.

Theorem 1. *Any public-coin distributed sketching protocol for computing a maximal matching with probability at least* 0.99 *must communicate* $\Omega\left(\frac{n^{1/2}}{e^{\Theta(\sqrt{\log n})}}\right)$ *bits from at least one player.*

We shall remark that the extension of Theorem 1 to the case when instead of at least one player, the *average* communication per player is $\Omega\left(\frac{\sqrt{n}}{e^{\Theta(\sqrt{\log n})}}\right)$ is standard. Basically, one needs to provide the "hard" input of the vertex communicating a large message to every vertex of the graph with constant probability and use the fact that simultaneous protocol cannot distinguish these two cases. We omit the details and instead refer the reader to [50, Section 3].

In the following, we first present our hard distribution for distributed sketching algorithms of maximal matching and then use it to prove a lower bound on sketch sizes and prove Theorem 1.

3.1 A Hard Distribution for Maximal Matching

Let $N \in \mathbb{N}$ be sufficiently large and consider the the distribution $\mathcal{D}_{\mathsf{MM}}$ on graphs with $n := n(N)$ vertices given below.

Distribution $\mathcal{D}_{\mathsf{MM}}$:
Parameters: $r = \frac{N}{e^{\Theta(\sqrt{\log N})}}$, $t = \frac{N}{3}$, $k = t$, $n = N - 2r + k \cdot 2r$.

(1) Fix an (r, t)-RS graph G^{RS} with vertex set $[N]$ using Proposition 2.1. Let $M_1^{\mathsf{RS}}, \ldots, M_t^{\mathsf{RS}}$ be its induced matchings.
(2) Pick $j^\star \in [t]$ uniformly at random and define $V^\star$ as the set of $2r$ vertices incident on $M_{j^\star}^{\mathsf{RS}}$.
(3) For $i = 1$ to k *independently*:
  (a) Let G_i be obtained from G^{RS} by dropping each edge w.p. $1/2$ independently and keeping the remaining edges.
(4) Pick a random permutation σ of $[n]$ and use it to *relabel* the vertices of the G_i graphs:
  (a) Enumerate the $N - 2r$ vertices of G^{RS} not in $V^\star$ (from the one with the smallest label to the largest). Let v be the ℓ^{th} vertex in the enumeration. Relabel the k vertices corresponding to v in $G_1, \ldots, G_k$ by the *same* label $\sigma(\ell)$.
  (b) For $i = 1$ to k:
    Enumerate the $2r$ vertices of G_i corresponding to vertices in $V^\star$ in G^{RS} (from the one with the smallest label to the largest). Relabel the ℓ^{th} vertex in the enumeration by $\sigma(N - 2r + (i-1) \cdot 2r + \ell)$.
(5) Let $G = (V, E)$ be the union of the graphs $G_1, \ldots, G_k$. That is, $V = [n]$ and for $u, v \in V$, $(u, v) \in E$ if and only if there exists $i \in [k]$ such that (u, v) is in the edge set of G_i.

Figure 1 gives an illustration of this distribution. From the description of the distribution, it can be seen that we are dealing with two different types of vertices that need to be treated differently. We define these vertices as follows:

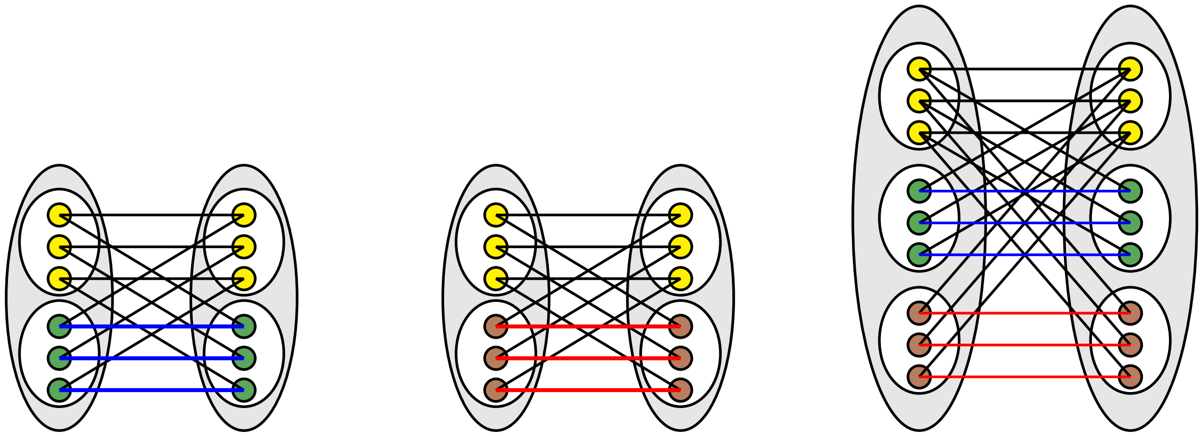
- **Public vertices:** The vertices with labels $\sigma(1), \ldots, \sigma(N-2r)$ are called *public* vertices. These are vertices that appear in *every* graph $G_1, \ldots, G_k$.
- **Unique vertices:** For any $i \in [k]$, the vertices with labels $\sigma(N - 2r + (i-1) \cdot 2r + 1), \ldots, \sigma(N - 2r + (i-1) \cdot 2r + 2r)$ are called *unique* vertices (of G_i). These vertices only appear in the graph G_i.

The very first step in the proof of Theorem 1 is the following claim regarding maximal matchings in graphs sampled from the distribution $\mathcal{D}_{\mathsf{MM}}$.

Claim 3.1. *With probability at least* $1 - 2^{-kr/10}$ *over the choice of graph* $G \sim \mathcal{D}_{\mathsf{MM}}$*, every maximal matching* M *of* G *has at least* $k \cdot r/4$ *edges whose both endpoints are unique vertices.*

Proof. For $i \in [k]$, let M_i be the matching in G_i corresponding to induced matching $M_{j^\star}^{\mathsf{RS}}$ in G^{RS}. Recall that the matchings $M_1, \ldots, M_k$ are on disjoint vertex sets and that $|M_i| \leq |M_{j^\star}^{\mathsf{RS}}| = r$. Also recall that each of the potential kr edges in $\cup_{i=1}^{k} M_i$ is removed with probability $1/2$, independently. Thus, $\mathbb{E}\left|\cup_{i=1}^{k} M_i\right| = k \cdot r/2$ and by Chernoff bound, the size of $\cup_{i=1}^{k} M_i$ is at least $k \cdot r/3$ with probability at least $1 - 2^{-kr/10}$. In the following, we condition on this event.

Suppose M is a maximal matching of G. Since there are $N - 2r$ public vertices, at most $N - 2r$ edges of M can have a public vertex



(a) Graph G_1 and its special matching M_1 (blue thick edges).

(b) Graph G_2 and its special matching M_2 (blue thick edges).

(c) Graph $G = G_1 \cup G_2$.

Figure 1: An illustration of the graphs in the hard distribution $\mathcal{D}_{\mathsf{MM}}$ for maximal matching. Each graph G_i is an RS graph with a "large" number of "large" induced matchings (for the purpose of this illustration, we have *not* removed half the edges of the RS graph randomly). Here, in the final graph G, the top blocks of vertices (yellow) of each G_i form the public vertices and the bottom block (green and brown) form the unique vertices (unlike this figure, the number of public vertices in graph G in general is much smaller than the total number of unique vertices).

as one endpoint. This leaves out at least

$$k \cdot r/3 - (N - 2r) > k \cdot r/3 - 3k \gg k \cdot r/4$$
$$\text{(for sufficiently large } N \text{ so } r = \frac{N}{e^{\Theta(\sqrt{\log N})}} > 36)$$

edges among $M_1, \ldots, M_k$ where both of their end points are unique and free to get matched by M. These edges must be in M, as M is maximal and since there are no additional edges in G supported on the end points of these edges (by the *induced* property of matchings in the RS graphs). This implies the claim as both endpoints of these edges are unique vertices. ∎

A Slight Change of The model: Public and Unique Players

Recall that in our model defined in Section 2, there is one player per every vertex of the graph. It turns out that for proving the lower bound, it is more convenient to consider the more general setting defined as follows. Instead of n players, we have $N - 2r + k \cdot N > n$ $(= N - 2r + k \cdot 2r)$ players partitioned into two groups, called *public* and *unique* players. There are in total $N - 2r$ public players denoted by $P := \{p_1, \ldots, p_{N-2r}\}$; each public player p_j gets all edges incident on the j^{th} public vertex in G (when the public vertices are enumerated from the one with the smallest label to the largest). We also have a set U of $k \cdot N$ unique players, consisting of a N players per each G_i, denoted by U_i. Each unique player $u_{i,j} \in U_i$ for $i \in [k]$ and $j \in [N]$ only sees the edges in G that correspond to edges incident on vertex j in G_i. Note that this implies that a unique player corresponding to a unique vertex u in G sees all the edges incident on vertex u in G (this is not the case for unique players that correspond to public vertices in G).

The only difference between this model and the original one is that there are now additionally k new "unique" copies of each public vertex, where the i^{th} copy can only see the edges of this vertex inside the graphs G_i. In our proof, we reveal to the referee for free the permutation σ and index $j^\star$ (we stress that σ and $j^\star$ are not revealed to the players), and hence also reveal the partitioning of vertices into public and unique vertices. As such, this new model can only be stronger than the old one for algorithms, as the referee can simply ignore the messages of unique players holding extra copies of the public vertices and run the protocol in the old model.

3.2 The Lower Bound for Distribution $\mathcal{D}_{\mathsf{MM}}$

We now prove the lower bound under this new model. Fix a *deterministic* protocol π for finding a maximal matching on graphs sampled from $\mathcal{D}_{\mathsf{MM}}$ with error probability at most 0.01. At the end, we will extend the lower bound to randomized algorithms on this distribution using an averaging argument (namely, the easy direction of Yao's minimax principle [53]).

We use $\Pi(P) := \pi(p_1), \ldots, \pi(p_{N-2r})$ to denote the collective messages of public players. For any $i \in [k]$, we further use $\Pi(U_i) := \pi(u_{i,1}), \ldots, \pi(u_{i,N})$ to denote the collective messages of unique players in G_i. Finally $\Pi(U) := \Pi(U_1), \ldots, \Pi(U_k)$ is the messages of all unique players and $\Pi := \Pi(P), \Pi(U)$ denotes all messages.

Let Σ, J be random variables representing the values of $\sigma, j^\star$ in the distribution $\mathcal{D}_{\mathsf{MM}}$. Let Π be a random variable representing the transcript of π (namely Π defined above). For $i \in [k]$ and $j \in [t]$, let $\mathsf{M}_{i,j} \in \{0,1\}^{M_j^{\mathsf{RS}}}$ be a vector-valued random variable that indicates for each edge e in the matchings in M_j^{RS} whether or not e was removed when constructing G_i. Namely, $\mathsf{M}_{i,j}(e) = 1$ if the edge e was not removed when constructing G_i, or *exists* in $\mathsf{M}_{i,j}$.

Recall that we assumed the referee is additionally provided with σ and $j^\star$ for free. Hence, the matching output by the referee, denoted by M_π, is a function of Π, σ and $j^\star$. We further write M_π^U

to denote the set of edges in M_π where both their endpoints are *unique* vertices. We use Claim 3.1 to lower bound the size of M^U_π.

Claim 3.2. $\mathbb{E}\left|M^U_\pi\right| \geq k \cdot r/5$.

Proof. With probability 0.01 the protocol errs, and with probability $1 - 2^{-rk/10} \gg 0.01$ the event in Claim 3.1 does not hold. By union bound, this means that with probability 0.98, the size of M^U_π should be at least $k \cdot r/4$, which implies the desired bound on the expectation. ∎

By Claim 3.2, M^U_π is rather large in expectation. We use this to argue that the messages of the players need to reveal a lot of information about the edges that exist in the graph, and in particular the edges corresponding to matchings between unique vertices, to enable the referee to output a large matching M^U_π. This is intuitive as the referee is outputting a large matching between the unique vertices and thus should know which edges exist to output them.

Lemma 3.3. $\mathbb{I}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} ; \Pi \mid \Sigma, \mathsf{J}) \geq k \cdot r/6$.

Proof. Firstly, note that edges of M^U_π all belong to $M^{\mathsf{RS}}_{j^\star}$ in the graphs $G_1, \ldots, G_k$, as both their endpoints are unique vertices. We use $\mathsf{M}_{out}(\pi, \sigma, j^\star) \subseteq \mathsf{M}_{1,j^\star}, \ldots, \mathsf{M}_{k,j^\star}$ to denote the random variables corresponding to edges in M^U_π (output by the referee) and $\overline{\mathsf{M}}_{out}$ to denote the remaining random variables among $\mathsf{M}_{1,j^\star}, \ldots, \mathsf{M}_{k,j^\star}$ (throughout this proof, we only focus on edges between unique vertices captured in $\mathsf{M}_{1,j^\star}, \ldots, \mathsf{M}_{k,j^\star}$).

By definition of mutual information,

$$\begin{aligned}
&\mathbb{I}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} ; \Pi \mid \Sigma, \mathsf{J}) \\
&\quad = \mathbb{H}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} \mid \Sigma, \mathsf{J}) - \mathbb{H}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} \mid \Pi, \Sigma, \mathsf{J}) \\
&\quad = k \cdot r - \mathbb{H}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} \mid \Pi, \Sigma, \mathsf{J}), \qquad (1)
\end{aligned}$$

as conditioned on Σ, J (but not Π), $\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}}$ is uniform over its support, which has size 2^{rk}, and thus we get the equality by Fact 2.2-(1). Our goal is now to upper bound the RHS of Eq (1).

Define $\mathsf{O} \in \{0, 1\}$ which is 1 if and only if the output of the protocol is correct. By applying chain rule of entropy (Fact 2.2-(4)) and since $\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} = \mathsf{M}_{out}, \overline{\mathsf{M}}_{out}$, we have,

$$\begin{aligned}
&\mathbb{H}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} \mid \Pi, \Sigma, \mathsf{J}) \\
&\quad \leq \mathbb{H}(\mathsf{M}_{out}, \overline{\mathsf{M}}_{out} \mid \mathsf{O}, \Pi, \Sigma, \mathsf{J}) + \mathbb{H}(\mathsf{O}) \\
&\quad \leq \mathbb{H}(\mathsf{M}_{out} \mid \mathsf{O}, \Pi, \Sigma, \mathsf{J}) + \mathbb{H}(\overline{\mathsf{M}}_{out} \mid \mathsf{M}_{out}, \mathsf{O}, \Pi, \Sigma, \mathsf{J}) + 1, \qquad (2)
\end{aligned}$$

as $\mathbb{H}(\mathsf{O}) \leq 1$ (by Fact 2.2-(1)). We now bound each of the remaining terms separately.

For the first term of Eq (2),

$$\begin{aligned}
\mathbb{H}(\mathsf{M}_{out} \mid \mathsf{O}, \Pi, \Sigma, \mathsf{J}) &= \Pr(\mathsf{O} = 0) \cdot \mathbb{H}(\mathsf{M}_{out} \mid \mathsf{O} = 0, \Pi, \Sigma, \mathsf{J}) \\
&\quad + \Pr(\mathsf{O} = 1) \cdot \mathbb{H}(\mathsf{M}_{out} \mid \mathsf{O} = 1, \Pi, \Sigma, \mathsf{J}) \\
&\leq \Pr(\mathsf{O} = 0) \cdot k \cdot r \leq k \cdot r/100,
\end{aligned}$$

where we used the fact that M_{out} has support 2^{rk} (and Fact 2.2-(1)), and that conditioned on $\mathsf{O} = 1$ and Π, Σ, J, entropy of M_{out} is zero because in this case, the correctness of the protocol (by conditioning $\mathsf{O} = 1$) ensures that all edges in M^U_π belong to the graph.

For the second term of Eq (2),

$$\begin{aligned}
&\mathbb{H}(\overline{\mathsf{M}}_{out} \mid \mathsf{M}_{out}, \mathsf{O}, \Pi, \Sigma, \mathsf{J}) \\
&\quad \leq \mathbb{H}(\overline{\mathsf{M}}_{out} \mid \Pi, \Sigma, \mathsf{J}) \\
&\quad\quad \text{(conditioning can only decrease entropy, Fact 2.2-(3))} \\
&\quad = \mathop{\mathbb{E}}_{\Pi, \sigma, j^\star} \left[\mathbb{H}(\overline{\mathsf{M}}_{out} \mid \Pi = \Pi, \Sigma = \sigma, \mathsf{J} = j^\star)\right] \\
&\quad = \mathop{\mathbb{E}}_{\Pi, \sigma, j^\star} \left[\log\left(\mathrm{supp}(\overline{\mathsf{M}}_{out} \mid \Pi = \Pi, \Sigma = \sigma, \mathsf{J} = j^\star)\right)\right] \\
&\quad\quad \text{(by Fact 2.2-(1))} \\
&\quad = \mathop{\mathbb{E}}_{\Pi, \sigma, j^\star} \left[k \cdot r - \left|M^U_\pi(\Pi, \sigma, j^\star)\right|\right] \\
&\quad\quad (\left|\overline{\mathsf{M}}_{out}\right| = k \cdot r - |\mathsf{M}_{out}|) \\
&\quad = k \cdot r - \mathbb{E}\left|M^U_\pi\right| \leq \frac{4}{5} \cdot k \cdot r. \qquad \text{(by Claim 3.2)}
\end{aligned}$$

Plugging in these bounds in Eq (2) and in Eq (1), we obtain that,

$$\begin{aligned}
\mathbb{I}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} ; \Pi \mid \Sigma, \mathsf{J}) &\geq k \cdot r - \left(k \cdot r/100 + \frac{4}{5} \cdot k \cdot r + 1\right) \\
&\geq k \cdot r/6,
\end{aligned}$$

concluding the proof. ∎

Our goal is now to upper bound $\mathbb{I}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} ; \Pi \mid \Sigma, \mathsf{J})$, the information about $\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}}$ revealed to the referee. The next lemma bounds this information by decomposing it to the information revealed by the public players P, and the sum of the informations revealed by each group U_i of unique players about their matching $\mathsf{M}_{i,\mathsf{J}}$. Intuitively, this can be done as the inputs of unique players from different G_i's are independent of each other (these inputs are only functions of which edges exists from G^{RS} in each G_i). As a result, the messages communicated by unique players inside one graph do not give extra information about another graph.

Lemma 3.4. *We have,*

$$\mathbb{I}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} ; \Pi \mid \Sigma, \mathsf{J}) \leq \mathbb{H}(\Pi(P)) + \sum_{i=1}^{k} \mathbb{I}(\mathsf{M}_{i,\mathsf{J}} ; \Pi(U_i) \mid \Sigma, \mathsf{J}).$$

Proof. Firstly, by chain rule of mutual information (Fact 2.2-(5)) and since $\Pi = \Pi(P), \Pi(U)$,

$$\begin{aligned}
&\mathbb{I}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} ; \Pi \mid \Sigma, \mathsf{J}) \\
&\quad = \mathbb{I}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} ; \Pi(U) \mid \Sigma, \mathsf{J}) \\
&\quad\quad + \mathbb{I}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} ; \Pi(P) \mid \Pi(U), \Sigma, \mathsf{J}) \\
&\quad \leq \mathbb{I}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} ; \Pi(U) \mid \Sigma, \mathsf{J}) + \mathbb{H}(\Pi(P)). \qquad (3)
\end{aligned}$$

We thus only need to upper bound the first term above.

Recall $\Pi(U) = \Pi(U_1), \ldots, \Pi(U_k)$. For $i \in [k]$, denote $\Pi(U^{<i}) = \Pi(U_1), \ldots, \Pi(U_{i-1})$. By chain rule (Fact 2.2-(5)),

$$\begin{aligned}
&\mathbb{I}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} ; \Pi(U) \mid \Sigma, \mathsf{J}) \\
&\quad = \sum_{i=1}^{k} \mathbb{I}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} ; \Pi(U_i) \mid \Pi(U^{<i}), \Sigma, \mathsf{J}).
\end{aligned}$$

We first show that for each $i \in [k]$,

$$\begin{aligned}
&\mathbb{I}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} ; \Pi(U_i) \mid \Pi(U^{<i}), \Sigma, \mathsf{J}) \\
&\quad \leq \mathbb{I}(\mathsf{M}_{1,\mathsf{J}}, \ldots, \mathsf{M}_{k,\mathsf{J}} ; \Pi(U_i) \mid \Sigma, \mathsf{J}), \qquad (4)
\end{aligned}$$

i.e., "dropping" the conditioning on $\Pi(U^{<i})$ only increases the information. This is because, after conditioning on Σ and J and any subset of $\{\mathsf{M}_{1,\mathsf{J}},\ldots,\mathsf{M}_{k,\mathsf{J}}\}$, the input of $u_{i,j}$ only depends on the (remaining) random coins used for deciding which edges of G^{RS} to remove to obtain G_i. Since G_i is constructed independently from all other $G_{i'}$, we get that the inputs of the *unique* players $u_{i,j}$ and $u_{i',j'}$ are independent of each other, for every $i' \neq i$ (we emphasize that this is *after* conditioning on σ and by input we mean which edges exist from G^{RS}). This also implies that $\Pi(U_i) \perp \Pi(U^{<i}) \mid \mathsf{M}_{1,\mathsf{J}},\ldots,\mathsf{M}_{k,\mathsf{J}},\Sigma,\mathsf{J}$, as $\Pi(U_i)$ and $\Pi(U^{<i})$ are deterministic functions of unique players' inputs. Hence, we can apply Proposition 2.4.

Denote $\mathsf{M}_{-i,\mathsf{J}} = \mathsf{M}_{1,\mathsf{J}},\ldots,\mathsf{M}_{i-1,\mathsf{J}},\mathsf{M}_{i+1,\mathsf{J}},\ldots,\mathsf{M}_{k,\mathsf{J}}$. By chain rule of mutual information,

$$\begin{aligned}&\mathbb{I}(\mathsf{M}_{1,\mathsf{J}},\ldots,\mathsf{M}_{k,\mathsf{J}}\,;\Pi(U_i)\mid\Sigma,\mathsf{J})\\&\quad=\mathbb{I}(\mathsf{M}_{i,\mathsf{J}}\,;\Pi(U_i)\mid\Sigma,\mathsf{J})+\mathbb{I}(\mathsf{M}_{-i,\mathsf{J}}\,;\Pi(U_i)\mid\mathsf{M}_{i,\mathsf{J}},\Sigma,\mathsf{J})\\&\quad=\mathbb{I}(\mathsf{M}_{i,\mathsf{J}}\,;\Pi(U_i)\mid\Sigma,\mathsf{J}),\end{aligned}$$

since $\mathbb{I}(\mathsf{M}_{-i,\mathsf{J}}\,;\Pi(U_i)\mid\mathsf{M}_{i,\mathsf{J}},\Sigma,\mathsf{J})=0$, as $\Pi(U_i)\perp\mathsf{M}_{-i,\mathsf{J}}\mid\mathsf{M}_{i,\mathsf{J}},\Sigma,\mathsf{J}$. The lemma now follows from this and Eq (3), Eq (4). ∎

Lemma 3.4 upper bounds the contribution of *public* players to revealing information about $\mathsf{M}_{1,\mathsf{J}},\ldots,\mathsf{M}_{k,\mathsf{J}}$ simply by the length (entropy) of their entire message. While quite a weak upper bound, this seems unavoidable as public players have a "good knowledge" of which edges of the graph are important and thus can directly inform the referee about those edges.

On the other hand, we now prove that, unlike public players, unique players in each U_i cannot reveal much information about their matchings without communicating much larger messages (by a factor of t, i.e., the *total* number of induced matchings in G^{RS}). This is established via a direct sum style argument which argues that since the players in U_i are *collectively* unaware of the identity of matching $\mathsf{M}_{i,\mathsf{J}}$, they need to reveal enough information about *every* induced matching in G_i in order to reveal enough information about the (unknown) matching $\mathsf{M}_{i,\mathsf{J}}$.

Lemma 3.5. *For any* $i\in[k]$, $\mathbb{I}(\mathsf{M}_{i,\mathsf{J}}\,;\Pi(U_i)\mid\Sigma,\mathsf{J})\leq\frac{1}{t}\cdot\mathbb{H}(\Pi(U_i))$.

Proof. Denote by Σ_i be the random variable representing the (partial) labeling function that was used by the algorithm for sampling from $\mathcal{D}_{\mathrm{MM}}$ to relabel the vertices of the graph G_i. Formally, Σ_i is the restriction of the permutation $\Sigma:[n]\to[n]$ to the domain $S_i=[N-2r]\cup\{N-2r+(i-1)\cdot 2r+1,\ldots,N-2r+i\cdot 2r\}$. Denote by Σ_{-i} the random variable representing the restriction of Σ to the domain $[n]\setminus S_i$. We identify Σ with (Σ_i,Σ_{-i}).

The input to players U_i (and consequently the message $\Pi(U_i)$) is uniquely determined by the matchings $\mathsf{M}_{i,1},\ldots,\mathsf{M}_{i,t}$ and the labeling function Σ_i, as these fully determine the graph G_i. Therefore, $\Pi(U_i)\perp\Sigma_{-i}\mid\mathsf{M}_{i,\mathsf{J}},\Sigma_i,\mathsf{J}$. By Proposition 2.4, it holds that

$$\mathbb{I}(\mathsf{M}_{i,\mathsf{J}}\,;\Pi(U_i)\mid\Sigma,\mathsf{J})\leq\mathbb{I}(\mathsf{M}_{i,\mathsf{J}}\,;\Pi(U_i)\mid\Sigma_i,\mathsf{J}).$$

We bound the the RHS of the above equation as follows,

$$\begin{aligned}\mathbb{I}(\mathsf{M}_{i,\mathsf{J}}\,;\Pi(U_i)\mid\Sigma_i,\mathsf{J})&=\mathop{\mathbb{E}}_{j\sim\mathsf{J}}\left[\mathbb{I}(\mathsf{M}_{i,j}\,;\Pi(U_i)\mid\Sigma_i,\mathsf{J}=j)\right]\\&=\frac{1}{t}\cdot\sum_{j=1}^{t}\mathbb{I}(\mathsf{M}_{i,j}\,;\Pi(U_i)\mid\Sigma_i),\end{aligned}$$

where the second equality is as the distribution of $(\mathsf{M}_{i,j},\Pi(U_i),\Sigma_i)$ is independent of the event $\mathsf{J}=j$ (in an informal sense, the unique players in U_i are unaware of which matching in the graph G_i is special even if they can all see the input of each other as well).

Since $(\mathsf{M}_{i,j+1},\ldots,\mathsf{M}_{i,t})\perp\mathsf{M}_{i,j}\mid\Sigma_i$ and by Proposition 2.3,

$$\begin{aligned}&\frac{1}{t}\cdot\sum_{j=1}^{t}\mathbb{I}(\mathsf{M}_{i,j}\,;\Pi(U_i)\mid\Sigma_i)\\&\qquad\leq\frac{1}{t}\cdot\sum_{j=1}^{t}\mathbb{I}(\mathsf{M}_{i,j}\,;\Pi(U_i)\mid\Sigma_i,\mathsf{M}_{i,j+1},\ldots,\mathsf{M}_{i,t}).\end{aligned}$$

By the chain rule of mutual information (Fact 2.2-(5)), the right hand side term simplifies to

$$\frac{1}{t}\cdot\mathbb{I}(\mathsf{M}_{i,1},\ldots,\mathsf{M}_{i,t}\,;\Pi(U_i)\mid\Sigma_i)\leq\frac{1}{t}\cdot\mathbb{H}(\Pi(U_i)),$$

finalizing the proof. ∎

Proof of Theorem 1. Let π be any protocol (deterministic or randomized) for the maximal matching problem over the distribution $\mathcal{D}_{\mathrm{MM}}$. By an averaging argument, we can fix the randomness of the protocol and obtain a deterministic protocol with the same worst-case length messages and probability of success. Fix such a protocol in the following and assume every player communicates b bits to the referee in the worst-case.

By combining Lemma 3.3, Lemma 3.4, and Lemma 3.5, and since $k=t$, we obtain that,

$$\begin{aligned}k\cdot r/6&\leq\mathbb{I}(\mathsf{M}_{1,\mathsf{J}},\ldots,\mathsf{M}_{k,\mathsf{J}}\,;\Pi\mid\Sigma,\mathsf{J})\leq\mathbb{H}(\Pi(P))+\frac{1}{t}\cdot\sum_{i=1}^{k}\mathbb{H}(\Pi(U_i))\\&\leq|P|\cdot b+\frac{kN\cdot b}{t}\leq Nb+\frac{k}{t}\cdot Nb=2Nb.\end{aligned}$$

Hence, we should have $2Nb\geq kr/6$ and so (since $k=t=N/3$),

$$b\geq\frac{1}{12N}\cdot kr=\frac{1}{12N}\cdot\frac{N}{3}\cdot r=\frac{r}{36}=\frac{N}{e^{\Theta(\sqrt{\log N})}}.$$

The total number of vertices, n, in the graph G, satisfies $n\geq N$ and $n\leq kN=N^2/3$, and hence $N=\Theta(\sqrt{n})$. This implies that the per-player communication cost has to be at least

$$b=\Omega\left(\frac{\sqrt{n}}{e^{\Theta(\sqrt{\log n})}}\right),$$

finalizing the proof of Theorem 1. ∎

We conclude this section by making the following remark that summarizes some key aspects of this lower bound.

Remark 3.6. *The lower bound in distribution* $\mathcal{D}_{\mathrm{MM}}$ *proven in this section holds even under* all *the following conditions:*

(i) The base graph G^{RS} *is known by all players and the referee (before dropping the edges);*
(ii) The choice of $j^{\star}$ *and* σ *is known to the referee (not the players);*
(iii) Public vertices know that they are public and additionally know the identity of all other public vertices;
(iv) The referee only needs to output a matching of size $k\cdot r/4$ *between the unique vertices (even if it is not maximal).*

Remark 3.6 follows directly from the proof of Theorem 1 in this section. We will use these properties to establish our lower bound for maximal independent set problem in the next section.

4 A Lower Bound for Maximal Independent Set

We now use a reduction from Theorem 1 to prove the following theorem.

Theorem 2. *Any public-coin distributed sketching protocol for computing a maximal independent set with probability at least* 0.99 *must communicate* $\Omega\left(\frac{n^{1/2}}{e^{\Theta(\sqrt{\log n})}}\right)$ *bits from at least one player.*

We prove Theorem 2 using a reduction from our lower bound in Theorem 1. We shall note that we are *not* giving a complete reduction from maximal matching to maximal independent set in the distributed sketching model. Our reduction crucially uses various properties of the hard distribution for Theorem 1, stated in Remark 3.6, and thus act as a reduction only for such instances. We are not aware of any general reduction between the two problems in the distributed sketching model (known reductions through line graphs in the LOCAL model are infeasible in this model as they would blow up communication complexity of protocols drastically).

A Reduction From Maximal Matching on Distribution $\mathcal{D}_{\mathsf{MM}}$

We design a reduction that given a graph $G \sim \mathcal{D}_{\mathsf{MM}}$, turns it into a graph H and uses a protocol for maximal independent set on H to find a maximal matching in G (or rather a large matching between unique vertices of G). To continue we need a definition.

Recall that in $\mathcal{D}_{\mathsf{MM}}$, each graph $G_1, \ldots, G_k$ was a copy of a base RS graph G^{RS} with edges dropped randomly with probability 1/2. For any $i \in [k]$, we define $M^{\mathsf{RS}}_{i,j^\star}$ to be a matching on vertices of G_i that is a copy of the $j^{\star\text{th}}$ induced matching of G^{RS} *before* dropping its edges in G_i randomly (hence, $M^{\mathsf{RS}}_{i,j^\star}$ is a superset of the induced matching of G_i). Also note that $M^{\mathsf{RS}}_{i,j^\star}$ for every $i \in [k]$ is supported on unique vertices. By construction, $M^{\mathsf{RS}}_{i,j^\star}$ is only a function of σ and $j^\star$ (to determine which matching to pick, and which vertices in G are endpoints of this matching).

We are now ready to give our reduction. Figure 2 gives an illustration of this reduction.

Reduction from maximal matching on $\mathcal{D}_{\mathsf{MM}}$:

(1) Suppose G is an n-vertex graph sampled from $\mathcal{D}_{\mathsf{MM}}$. The players collectively create the graph H on $2n$ vertices as follows:
  (a) Each vertex $u \in G$ creates two copies u^ℓ and u^r of the same vertex, and connect u^ℓ to v^ℓ and u^r to v^r for every neighbor v of u in G. This step creates two identical copies of G on two disjoint sets of vertices denoted by V^ℓ and V^r.
  (b) Each *public* vertex u in G adds an edge between u^ℓ and v^r, and also between u^r and v^ℓ, for every *public* vertex v in G (by Remark 3.6 we assume public vertices know identity of other public vertices). Let H be this new graph.

(2) The players run the distributed sketching protocol for maximal independent set on H by each vertex u simulating the protocol for both vertices u^ℓ and u^r and sending their messages to the referee. The referee computes the maximal independent set $\mathcal{M}$ of H.

(3) The referee computes the matchings $M^{\mathsf{RS}}_{i,j^\star}$ for every $i \in [k]$ (by Remark 3.6, referee knows $(\sigma, j^\star)$ and can construct this matching). Then, the referee creates two matchings M^ℓ and M^r as follows: for any pair of vertices $(u,v) \in M^{\mathsf{RS}}_{i,j^\star}$ for $i \in [k]$, if u^ℓ, v^ℓ (resp., u^r, v^r) are *not* both in $\mathcal{M}$, add an edge (u^ℓ, v^ℓ) to M^ℓ (resp. (u^r, v^r) to M^r).

(4) If $|M^\ell| \geq |M^r|$, the referee outputs the pre-image of edges of M^ℓ in G as the final matching (that is, for every $(u^\ell, v^\ell) \in M^\ell$, the final matching contains the edge (u,v)). Otherwise, the referee outputs the pre-image of the edges of M^r.

Similar to Section 3, we use P^ℓ, P^r, and U^ℓ, U^r, to denote the copies of public vertices and unique vertices of G in H, respectively. We prove the lower bound by showing that the matching output by the reduction is a valid matching of size at least $k \cdot r/4$ in G between unique vertices, and apply the last part of Remark 3.6 to conclude the lower bound. The main step of the proof is the following lemma that establishes the correctness of the reduction.

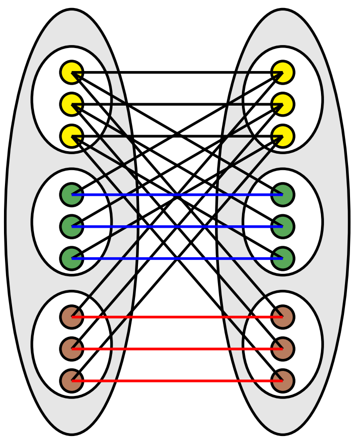
Lemma 4.1. *Suppose S is any maximal independent set in H such that $S \cap P^\ell = \emptyset$ (resp. $S \cap P^r = \emptyset$). Let (u,v) be any edge in any $M^{\mathsf{RS}}_{i,j^\star}$ for $i \in [k]$. Then (u,v) survived the random sampling (in $\mathcal{D}_{\mathsf{MM}}$) in G if and only if* not *both of u^ℓ, v^ℓ belong to S (resp. not both of u^r, v^r belong to S).*

Proof. We only prove the lemma for P^ℓ; the case for P^r follows by symmetry.

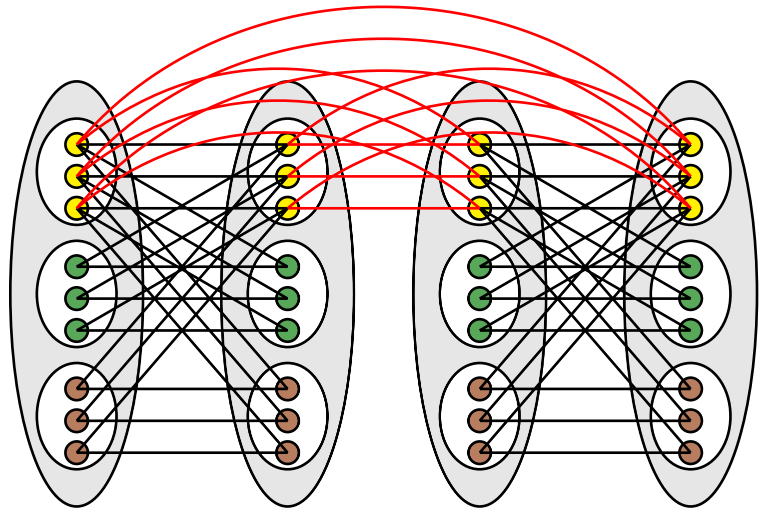
Since S is an independent set in H, there can be no edge between u^ℓ, v^ℓ if they both belong to S, and hence their pre-image u, v cannot have an edge in G. As such, (u,v) has not survived the random sampling, proving the first direction of the lemma.

Now consider any pair of vertices u^ℓ, v^ℓ where the edge (u,v) has not survived the random sampling in G. Since P^ℓ has no intersection with S and vertices in U^ℓ have no edges to P^r, U^r, the maximality of S ensures that $S \cap U^\ell$ is a maximal independent set on the induced subgraph on U^ℓ. However, the induced subgraph of U^ℓ is the collection of induced matchings of G_i's and hence the only possible edge incident on at least one of the vertices u^ℓ, v^ℓ is the potential edge (u^ℓ, v^ℓ). As (u,v) has not survived the random sampling, (u^ℓ, v^ℓ) does not exists in H, and thus, by maximality of S, both u^ℓ, v^ℓ should be part of S (as no edges are incident on neither of them). ∎

Proof of Theorem 2. Let π be any protocol (deterministic or randomized) for maximal independent set and let b denote the worst-case length of messages communicated by any player.



(a) Graph G from $\mathcal{D}_{\mathsf{MM}}$



(b) Graph H in the reduction.

Figure 2: An illustration of the graphs created by the reduction for the maximal independent set problem given a graph $G \sim \mathcal{D}_{\mathsf{MM}}$ as input. The reduction involves creating two identical copies of G and then connecting all public vertices together (red edges).

As explained in the reduction, each vertex $u \in G$ can create the neighborhood of both u^ℓ and v^ℓ correctly in H, and thus simulate π for them in H consistently with at most $2 \cdot b$ communication from u. By definition, π outputs a correct $\mathcal{M}$ with probability at least 0.99. Whenever this happens, by construction of H, we know that at least one of $\mathcal{M} \cap P^\ell$ or $\mathcal{M} \cap P^r$ should be empty (since all vertices in P^ℓ and P^r are connected to each other). Conditioned on this event, by Lemma 4.1, at least one of M^ℓ or M^r contains *all* edges between unique vertices U^ℓ or U^r, and thus the referee recovers the entire matching between unique vertices in G.

By Remark 3.6, the lower bound of Theorem 1 implies that $2b = \Omega(n^{1/2}/e^{\Theta(\sqrt{\log n})})$ which concludes the proof. ∎

Acknowledgements

Sepehr Assadi would like to thank Michael Kapralov, Sanjeev Khanna, Omri Weinstein, and Huacheng Yu for helpful discussions. We are also grateful to the anonymous reviewers of PODC 2020 for invaluable suggestions and comments on the related work and the presentation of this paper.

References

[1] Kook Jin Ahn, Sudipto Guha, and Andrew McGregor. 2012. Analyzing Graph Structure via Linear Measurements. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '12)*. SIAM, 459–467. http://dl.acm.org/citation.cfm?id=2095116.2095156 1, 2, 3

[2] Kook Jin Ahn, Sudipto Guha, and Andrew McGregor. 2012. Graph sketches: sparsification, spanners, and subgraphs. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2012, Scottsdale, AZ, USA, May 20-24, 2012*. 5–14. https://doi.org/10.1145/2213556.2213560 1, 3

[3] Kook Jin Ahn, Sudipto Guha, and Andrew McGregor. 2013. Spectral Sparsification in Dynamic Graph Streams. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 16th International Workshop, APPROX 2013, and 17th International Workshop, RANDOM 2013, Berkeley, CA, USA, August 21-23, 2013. Proceedings*. 1–10. 1

[4] Noga Alon. 2002. Testing subgraphs in large graphs. *Random Struct. Algorithms* 21, 3-4 (2002), 359–370. 3

[5] Noga Alon, Ankur Moitra, and Benny Sudakov. 2012. 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*. 1079–1090. 3

[6] Noga Alon, Noam Nisan, Ran Raz, and Omri Weinstein. 2015. Welfare Maximization with Limited Interaction. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*. 1499–1512. 3

[7] Noga Alon and Asaf Shapira. 2006. A Characterization of Easily Testable Induced Subgraphs. *Combinatorics, Probability & Computing* 15, 6 (2006), 791–805. 3

[8] Alexandr Andoni, Assaf Goldberger, Andrew McGregor, and Ely Porat. 2013. Homomorphic fingerprints under misalignments: sketching edit and shift distances. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*. 931–940. 2

[9] Sepehr Assadi. 2017. Combinatorial Auctions Do Need Modest Interaction. In *Proceedings of the 2017 ACM Conference on Economics and Computation, EC '17, Cambridge, MA, USA, June 26-30, 2017*. 145–162. 3

[10] Sepehr Assadi and Aaron Bernstein. 2019. 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*. 11:1–11:20. 3

[11] Sepehr Assadi, Yu Chen, and Sanjeev Khanna. 2019. Sublinear Algorithms for $(\Delta + 1)$ Vertex Coloring. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*. 767–786. 1, 2

[12] Sepehr Assadi, Sanjeev Khanna, and Yang Li. 2016. Tight bounds for single-pass streaming complexity of the set cover problem. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*. 698–711. 3

[13] Sepehr Assadi, Sanjeev Khanna, and Yang Li. 2017. On Estimating Maximum Matching Size in Graph Streams. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*. 1723–1742. 3

[14] Sepehr Assadi, Sanjeev Khanna, Yang Li, and Grigory Yaroslavtsev. 2016. Maximum Matchings in Dynamic Graph Streams and the Simultaneous Communication Model. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*. 1345–1364. 2, 3

[15] László Babai, Anna Gál, Peter G. Kimmel, and Satyanarayana V. Lokam. 2003. Communication Complexity of Simultaneous Messages. *SIAM J. Comput.* 33, 1 (2003), 137–166. 3

[16] Florent Becker, Adrian Kosowski, Martín Matamala, Nicolas Nisse, Ivan Rapaport, Karol Suchan, and Ioan Todinca. 2015. Allowing each node to communicate only once in a distributed system: shared whiteboard models. *Distributed Comput.* 28, 3 (2015), 189–200. 2

[17] Florent Becker, Martín Matamala, Nicolas Nisse, Ivan Rapaport, Karol Suchan, and Ioan Todinca. 2011. Adding a Referee to an Interconnection Network: What Can(not) Be Computed in One Round. In *25th IEEE International Symposium on Parallel and Distributed Processing, IPDPS 2011, Anchorage, Alaska, USA, 16-20 May, 2011 - Conference Proceedings.* 508–514. 2, 3

[18] Florent Becker, Pedro Montealegre, Ivan Rapaport, and Ioan Todinca. 2014. The Simultaneous Number-in-Hand Communication Model for Networks: Private Coins, Public Coins and Determinism. In *Structural Information and Communication Complexity - 21st International Colloquium, SIROCCO 2014, Takayama, Japan, July 23-25, 2014. Proceedings.* 83–95. 2

[19] Florent Becker, Pedro Montealegre, Ivan Rapaport, and Ioan Todinca. 2018. The Impact of Locality on the Detection of Cycles in the Broadcast Congested Clique Model. In *LATIN 2018: Theoretical Informatics - 13th Latin American Symposium, Buenos Aires, Argentina, April 16-19, 2018, Proceedings.* 134–145. 2

[20] Felix A Behrend. 1946. On sets of integers which contain no three terms in arithmetical progression. *Proceedings of the National Academy of Sciences of the United States of America* 32, 12 (1946), 331. 3

[21] Bertinoro. Bertinoro workshop 2014, problem 65. https://sublinear.info/65. (????). Accessed: 2020-2-14. 2

[22] Sayan Bhattacharya, Monika Henzinger, Danupon Nanongkai, and Charalampos E. Tsourakakis. 2015. Space- and Time-Efficient Algorithm for Maintaining Dense Subgraphs on One-Pass Dynamic Streams. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015.* 173–182. 1

[23] Yitzhak Birk, Nathan Linial, and Roy Meshulam. 1993. On the uniform-traffic capacity of single-hop interconnections employing shared directional multichannels. *IEEE Transactions on Information Theory* 39, 1 (1993), 186–191. 3

[24] Mark Braverman and Rotem Oshman. 2017. A Rounds vs. Communication Tradeoff for Multi-Party Set Disjointness. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017.* 144–155. 3

[25] Ashok K. Chandra, Merrick L. Furst, and Richard J. Lipton. 1983. Multi-Party Protocols. In *Proceedings of the 15th Annual ACM Symposium on Theory of Computing, 25-27 April, 1983, Boston, Massachusetts, USA.* 94–99. 3

[26] Graham Cormode, Jacques Dark, and Christian Konrad. 2019. Independent Sets in Vertex-Arrival Streams. In *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece.* 45:1–45:14. 2, 3

[27] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of information theory (2. ed.).* Wiley. 3

[28] Debarati Das, Michal Koucký, and Michael E. Saks. 2018. Lower Bounds for Combinatorial Algorithms for Boolean Matrix Multiplication. In *35th Symposium on Theoretical Aspects of Computer Science, STACS 2018, February 28 to March 3, 2018, Caen, France.* 23:1–23:14. 3

[29] Shahar Dobzinski, Noam Nisan, and Sigal Oren. 2014. Economic efficiency requires interaction. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014.* 233–242. 3

[30] Andrew Drucker, Fabian Kuhn, and Rotem Oshman. 2014. On the power of the congested clique model. In *ACM Symposium on Principles of Distributed Computing, PODC '14, Paris, France, July 15-18, 2014.* 367–376. 2, 3

[31] Martin Farach-Colton and Meng-Tsung Tsai. 2016. Tight Approximations of Degeneracy in Large Graphs. In *LATIN 2016: Theoretical Informatics - 12th Latin American Symposium, Ensenada, Mexico, April 11-15, 2016, Proceedings.* 429–440. 1

[32] Eldar Fischer, Eric Lehman, Ilan Newman, Sofya Raskhodnikova, Ronitt Rubinfeld, and Alex Samorodnitsky. 2002. Monotonicity testing over general poset domains. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada.* 474–483. 3

[33] Orr Fischer, Rotem Oshman, and Uri Zwick. 2016. Public vs. Private Randomness in Simultaneous Multi-party Communication Complexity. In *Structural Information and Communication Complexity - 23rd International Colloquium, SIROCCO 2016, Helsinki, Finland, July 19-21, 2016, Revised Selected Papers.* 60–74. 3

[34] Jacob Fox, Hao Huang, and Benny Sudakov. 2015. On graphs decomposable into induced matchings of linear sizes. *arXiv preprint arXiv:1512.07852* (2015). 3

[35] Mohsen Ghaffari, Themis Gouleakis, Christian Konrad, Slobodan Mitrovic, and Ronitt Rubinfeld. 2018. Improved Massively Parallel Computation Algorithms for MIS, Matching, and Vertex Cover. In *Proceedings of the 2018 ACM Symposium on Principles of Distributed Computing, PODC 2018, July 23-27, 2018.* 129–138. 2

[36] Ashish Goel, Michael Kapralov, and Sanjeev Khanna. 2012. On the Communication and Streaming Complexity of Maximum Bipartite Matching. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '12).* SIAM, 468–485. http://dl.acm.org/citation.cfm?id=2095116.2095157 3

[37] Sudipto Guha, Andrew McGregor, and David Tench. 2015. Vertex and Hyperedge Connectivity in Dynamic Graph Streams. In *Proceedings of the 34th ACM Symposium on Principles of Database Systems, PODS 2015, Melbourne, Victoria, Australia, May 31 - June 4, 2015.* 241–247. 1

[38] Johan Håstad and Avi Wigderson. 2003. Simple analysis of graph tests for linearity and PCP. *Random Struct. Algorithms* 22, 2 (2003), 139–160. 3

[39] Tomasz Jurdzinski, Krzysztof Lorys, and Krzysztof Nowicki. 2018. Communication Complexity in Vertex Partition Whiteboard Model. In *Structural Information and Communication Complexity - 25th International Colloquium, SIROCCO 2018, Ma'ale HaHamisha, Israel, June 18-21, 2018, Revised Selected Papers.* 264–279. 2, 3

[40] Tomasz Jurdzinski and Krzysztof Nowicki. 2017. Brief Announcement: On Connectivity in the Broadcast Congested Clique. In *31st International Symposium on Distributed Computing, DISC 2017, October 16-20, 2017, Vienna, Austria.* 54:1–54:4. 2

[41] Tomasz Jurdzinski and Krzysztof Nowicki. 2018. Connectivity and Minimum Cut Approximation in the Broadcast Congested Clique. In *Structural Information and Communication Complexity - 25th International Colloquium, SIROCCO 2018, Ma'ale HaHamisha, Israel, June 18-21, 2018, Revised Selected Papers.* 331–344. 2

[42] Michael Kapralov. 2013. 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.* 1679–1697. https://doi.org/10.1137/1.9781611973105.121 3

[43] Michael Kapralov, Yin Tat Lee, Cameron Musco, Christopher Musco, and Aaron Sidford. 2014. Single Pass Spectral Sparsification in Dynamic Streams. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014.* 561–570. https://doi.org/10.1109/FOCS.2014.66 1

[44] Michael Kapralov, Jelani Nelson, Jakub Pachocki, Zhengyu Wang, David P. Woodruff, and Mobin Yahyazadeh. 2017. Optimal Lower Bounds for Universal Relation, and for Samplers and Finding Duplicates in Streams. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017.* 475–486. 2

[45] Christian Konrad. 2015. Maximum Matching in Turnstile Streams. In *Algorithms - ESA 2015 - 23rd Annual European Symposium, September 14-16, 2015, Proceedings.* 840–852. 3

[46] Silvio Lattanzi, Benjamin Moseley, Siddharth Suri, and Sergei Vassilvitskii. 2011. Filtering: a method for solving graph problems in MapReduce. In *SPAA 2011: Proceedings of the 23rd Annual ACM Symposium on Parallelism in Algorithms and Architectures, San Jose, CA, USA, June 4-6, 2011 (Co-located with FCRC 2011).* 85–94. https://doi.org/10.1145/1989493.1989505 2

[47] Nati Linial, Toniann Pitassi, and Adi Shraibman. 2019. On the Communication Complexity of High-Dimensional Permutations. In *10th Innovations in Theoretical Computer Science Conference, ITCS 2019, January 10-12, 2019, San Diego, California, USA.* 54:1–54:20. 3

[48] Andrew McGregor, David Tench, Sofya Vorotnikova, and Hoa T. Vu. 2015. Densest Subgraph in Dynamic Graph Streams. In *Mathematical Foundations of Computer Science 2015 - 40th International Symposium, MFCS 2015, Milan, Italy, August 24-28, 2015, Proceedings, Part II.* 472–482. 1

[49] Pedro Montealegre, Sebastian Perez-Salazar, Ivan Rapaport, and Ioan Todinca. 2018. Two Rounds Are Enough for Reconstructing Any Graph (Class) in the Congested Clique Model. In *Structural Information and Communication Complexity - 25th International Colloquium, SIROCCO 2018, Ma'ale HaHamisha, Israel, June 18-21, 2018, Revised Selected Papers.* 134–148. 3

[50] Jelani Nelson and Huacheng Yu. 2019. Optimal Lower Bounds for Distributed and Streaming Spanning Forest Computation. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019.* 1844–1860. 2, 3, 4

[51] Imre Z Ruzsa and Endre Szemerédi. 1978. Triple systems with no six points carrying three triangles. *Combinatorics (Keszthely, 1976), Coll. Math. Soc. J. Bolyai* 18 (1978), 939–945. 3

[52] Terence Tao and Van H Vu. 2006. *Additive combinatorics.* Vol. 105. Cambridge University Press. 3

[53] Andrew Chi-Chih Yao. 1977. Probabilistic Computations: Toward a Unified Measure of Complexity (Extended Abstract). In *18th Annual Symposium on Foundations of Computer Science, Providence, Rhode Island, USA, 31 October - 1 November 1977.* 222–227. 5