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 send 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, admit efficient algorithms in this model that only require sketches of size \( \text{polylog}(n) \) 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 Rusza-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

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 \((u, v)\). 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.

\footnote{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 \( s_w := \sum_{z \in N(w)} (z \cdot n + w) - \sum_{z \in N(w) \cap N(v)} (w - n - z) \) 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 $(A + 1)$ 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 $n^{0(1)}$ 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-Szemeredi 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-Szemeredi 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-Szemeredi 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 $O(\log^2 n)$ 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 \( \text{shared} \) 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 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 \( \text{shared} \) 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 \( G^{RS}(V, E) \) 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 \( M_1^{RS}, \ldots, M_t^{RS} \), 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 \( r = \frac{N}{e^{N/\sqrt{\log(N)}}} \) 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, 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 texts by Cover and Thomas [27] for a broader introduction.

For random variables \( A \), \( B \), we use \( I(A) \) and \( I(A; B) \) 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. \( 0 \leq I(A) \leq \log |\text{supp}(A)| \) (where \( \text{supp}(A) \) denote the support of \( A \). The right equality holds iff \( A \) is uniformly distributed.
2. \( I(A; B) \geq 0 \). The equality holds iff \( A \) and \( B \) are independent.
3. Conditioning on a random variable reduces entropy: \( H(A \mid B, C) \leq H(A \mid B) \). The equality holds iff \( A \perp B \mid C \).
We will use the following two standard inequalities regarding the
entropy.

\[ H(A; B | C) = H(A | C) + H(B | C, A). \]

(5) Chain rule for mutual information:
\[ I(A; B; C | D) = I(A; C | D) + I(B; C | A, D). \]

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

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

**Proof.** By Fact 2.2-(3), since \( A \perp D | C \), we have \( \mathbb{H}(A | C) = \mathbb{H}(A | C, D) \) and since conditioning can only decrease the entropy, \( \mathbb{H}(A | B, C) \leq \mathbb{H}(A | C, D) \). As such,
\[
\mathbb{I}(A; B | C) = \mathbb{H}(A | C) - \mathbb{H}(A | C, B) \\
\leq \mathbb{H}(A | C, D) - \mathbb{H}(A | C, B, D) = \mathbb{I}(A; B | C, D),
\]
concluding the proof.

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

**Proof.** By Fact 2.2-(3), since \( A \perp D | B, C \), we have \( \mathbb{H}(A | B, C) = \mathbb{H}(A | B, C, D) \) and since conditioning can only reduce the entropy, \( \mathbb{H}(A | B) \geq \mathbb{H}(A | C, D) \). As such,
\[
\mathbb{I}(A; B | C) = \mathbb{H}(A | C) - \mathbb{H}(A | B, C) \\
\geq \mathbb{H}(A | D, C) - \mathbb{H}(A | B, C, D) = \mathbb{I}(A; B | C, D),
\]
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 \( \Omega \left( \frac{n^{1/2}}{\epsilon^{0.5 \sqrt{n}} \ln \frac{1}{\epsilon}} \right) \) 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 \( \Omega \left( \frac{n}{\epsilon^{0.5 \sqrt{n}} \ln \frac{1}{\epsilon}} \right) \) 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 distribution \( D_{MM} \) on graphs with \( n := n(N) \) vertices given below.

<table>
<thead>
<tr>
<th>Distribution ( D_{MM} ):</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parameters:</strong> ( r = \frac{N}{\epsilon^{0.5 \sqrt{n}} \ln \frac{1}{\epsilon}}, \quad t = \frac{N}{3}, \quad k = t, \quad n = N - 2r + k \cdot 2r. )</td>
</tr>
</tbody>
</table>

(1) Fix an \( (r, t) \)-RS graph \( G^{RS} \) with vertex set \( [N] \) using Proposition 2.1. Let \( M_i^{RS} \), \( i = 1, \ldots, t \), be its induced matchings.

(2) Pick \( j^* \in [t] \) uniformly at random and define \( V^* \) as the set of \( 2r \) vertices incident on \( M_{j^*}^{RS} \).

(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 \( \sigma \) 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 \( V^* \) (from the one with the smallest label to the largest). Let \( \nu \) be the \( \sigma(i) \)th vertex in the enumeration. Relabel the \( k \) vertices corresponding to \( \nu \) in \( G_1, \ldots, G_k \) by the same label \( \sigma(i) \).

(b) \( \nu_v \) is a hard distribution with the extra condition that for any \( i \leq k \), the \( k \) vertices corresponding to \( \nu \) in \( G_i \) are relabeled by the same label \( \sigma(i) \).

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 \( D_{MM} \).

**Claim 3.1.** With probability at least \( 1 - 2^{-kr/10} \) over the choice of graph \( G \sim D_{MM} \), 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 \( M_{j^*}^{RS} \) in \( G^{RS} \). Recall that the matchings \( M_1, \ldots, M_k \) are on disjoint vertex sets and that \( |M_i| \leq |M_{j^*}^{RS}| = r \).

Also recall that each of the potential \( kr \) edges in \( \bigcup_{i=1}^{k} M_i \) is removed with probability \( 1/2 \), independently. Thus, \( \mathbb{E} \left| \bigcup_{i=1}^{k} M_i \right| = k \cdot r/2 \) and by Chernoff bound, the size of \( \bigcup_{i=1}^{k} M_i \) 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.
We also have a set $U = \{u_1, \ldots, u_k\}$ 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 $N = 2r + k \cdot N > n \geq \sqrt{n - 2r + k \cdot 2r}$ 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 $U = \{u_1, \ldots, u_k\}$ unique players, consisting of $k$ 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 $u$ 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 $\pi$ and index $j^*$ (we stress that $\pi$ and $j^*$ 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 $D_{MM}$

We now prove the lower bound under this new model. Fix a deterministic protocol $\pi$ for finding a maximal matching on graphs sampled from $D_{MM}$ 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 $\Sigma, J$ be random variables representing the values of $\sigma, j^*$ in the distribution $D_{MM}$. Let $f_i$ be a random variable representing the transcript of $\pi$ (namely $\Pi$ defined above). For $i \in [k]$ and $j \in [t]$, let $M_{i,j} \in \{0,1\}^{M_{RS}^{ij}}$ be a vector-valued random variable that indicates for each edge $e$ in the matchings in $\mathcal{M}_{i,j}$ whether or not $e$ was removed when constructing $G_i$. Namely, $M_{i,j}(e) = 1$ if the edge $e$ was not removed when constructing $G_i$, or exists in $M_{i,j}$.

Recall that we assumed the referee is additionally provided with $\sigma$ and $j^*$ for free. Hence, the matching output by the referee, denoted by $\Pi_\sigma$, is a function of $\Pi, \sigma$ and $j^*$. We further write $\Pi_i$
To denote the set of edges in $M_*$ where both their endpoints are unique vertices. We use Claim 3.1 to lower bound the size of $M_U^*$.

Claim 3.2. $E|M_U^*| \geq k \cdot r/5$.

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

By Claim 3.2, $M_U^*$ 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 $M^*$. 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. $I(M_{i,1}, \ldots, M_{i,k}; \Pi | \Sigma, J) \geq k \cdot r/6$.

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

By definition of mutual information,

$$I(M_{i,1}, \ldots, M_{i,k}; \Pi | \Sigma, J) = E(H(M_{i,1}, \ldots, M_{i,k} | \Sigma, J) - H(M_{i,1}, \ldots, M_{i,k} | \Pi, \Sigma, J))$$

$$= k \cdot r - E(H(M_{i,1}, \ldots, M_{i,k} | \Pi, \Sigma, J))$$

as conditioned on $\Sigma, J$ (but not $\Pi$), $M_{i,1}, \ldots, M_{i,k}$ is uniform over its support, which has size $2^{r/2}$, and thus we get the equality by Fact 2.2-(1). Our goal is now to upper bound the RHS of Eq (1).

Define $\Omega \in \{0, 1\}$ 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 $M_{i,1}, \ldots, M_{i,k} = M_{\text{out}}, M_{\text{out}}$, we have,

$$E(H(M_{i,1}, \ldots, M_{i,k} | \Pi, \Sigma, J)) \leq E(H(M_{\text{out}}, M_{\text{out}}, \Omega, \Pi, \Sigma, J)) + H(\Omega)$$

$$\leq E(H(M_{\text{out}} | \Omega, \Pi, \Sigma, J) + H(M_{\text{out}}, \Omega, \Pi, \Sigma, J) + 1$$

(2)

as $H(\Omega) \leq 1$ (by Fact 2.2-(1)). We now bound each of the remaining terms separately.

For the first term of Eq (2),

$$E(H(M_{\text{out}} | \Omega, \Pi, \Sigma, J)) = \Pr(\Omega = 0) \cdot E(H(M_{\text{out}} | \Omega = 0, \Pi, \Sigma, J)$$

$$+ \Pr(\Omega = 1) \cdot E(H(M_{\text{out}} | \Omega = 1, \Pi, \Sigma, J) \leq \Pr(\Omega = 0) \cdot k \cdot r \leq k \cdot r/100,$$

where we used the fact that $M_{\text{out}}$ has support $2^{r/2}$ (and Fact 2.2-(1)), and that conditioned on $\Omega = 1$ and $\Pi, \Sigma, J$, entropy of $M_{\text{out}}$ is zero because in this case, the correctness of the protocol (by conditioning $\Omega = 1$) ensures that all edges in $M_{\text{out}}$ belong to the graph.

For the second term of Eq (2),

$$E(H(M_{\text{out}}, \Omega, \Pi, \Sigma, J) \leq E(H(M_{\text{out}} | \Pi, \Sigma, J)$$

(conditioning can only decrease entropy, Fact 2.2-(3))

$$= \sum_{\Pi, \sigma, j^*} E(H(M_{\text{out}} | \Pi = \Pi, \Sigma = \sigma, J = j^*))$$

$$= \sum_{\Pi, \sigma, j^*} \log \left(\supp(M_{\text{out}} | \Pi = \Pi, \Sigma = \sigma, J = j^*)\right)$$

(by Fact 2.2-(1))

$$= \sum_{\Pi, \sigma, j^*} k \cdot r - M_{\text{out}}^U(\Pi, \sigma, j^*)$$

$$\leq k \cdot r - E(M_{\text{out}}^U)$$

(by Claim 3.2)

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

$$I(M_{i,1}, \ldots, M_{i,k}; \Pi | \Sigma, J) \geq k \cdot r - \left(\frac{k \cdot r}{100} + \frac{4}{5} \cdot \frac{k \cdot r}{10} + 1\right)$$

$$\geq k \cdot r/6$$

concluding the proof.

Our goal is now to upper bound $I(M_{i,1}, \ldots, M_{i,k}; \Pi | \Sigma, J)$, the information about $M_{i,1}, \ldots, M_{i,k}$ revealed to the referee. The next lemma bounds this information by decomposing it to the information revealed by the public players $P_i$ and the sum of the informations revealed by each group $U_i$ of unique players about their matching $M_{i,1}$. 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 $G^R$ 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,

$$I(M_{i,1}, \ldots, M_{i,k}; \Pi | \Sigma, J) \leq E(H(P)) + \sum_{i=1}^{k} I(M_{i,j}; \Pi(U_i) | \Sigma, J).$$

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

$$I(M_{i,1}, \ldots, M_{i,k}; \Pi | \Sigma, J) = I(M_{i,1}, \ldots, M_{i,k}; U_i | \Sigma, J)$$

$$+ I(M_{i,1}, \ldots, M_{i,k}; P | U_i, \Sigma, J)$$

$$\leq I(M_{i,1}, \ldots, M_{i,k}; P | U_i, \Sigma, J)$$

$$\leq \sum_{i=1}^{k} I(M_{i,j}; \Pi(U_i) | \Sigma, J) + H(P_i).$$

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

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

$$I(M_{i,1}, \ldots, M_{i,k}; P | \Sigma, J)$$

$$= \sum_{i=1}^{k} I(M_{i,1}, \ldots, M_{i,k}; U_i | \Pi(U^{<i}), \Sigma, J).$$

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

$$I(M_{i,1}, \ldots, M_{i,k}; U_i | \Pi(U^{<i}), \Sigma, J) \leq I(M_{i,1}, \ldots, M_{i,k}; U_i | \Sigma, J),$$

(4)
i.e., “dropping” the conditioning on $\Pi(U^{<i})$ only increases the information. This is because, after conditioning on $\Sigma$ and $J$ and any subset of $\{M_{1,i}, \ldots, M_{k,i}\}$, 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 $G_j$, we get that the inputs of the unique players $u_{i,j}$ and $w_{t',j'}$ are independent of each other, for every $t' \neq i$ (we emphasize that this is after conditioning on $\sigma$ and by input we mean which edges exist from $G^{RS}$). This also implies that $\Pi(U_j) \perp \Pi(U^{<t}) \mid M_{1,i}, \ldots, M_{k,i} \Sigma, J$, as $\Pi(U_j)$ and $\Pi(U^{<t})$ are deterministic functions of unique players’ inputs. Hence, we can apply Proposition 2.4.

Denote $M_{-i,j} = M_{1,j}, \ldots, M_{i-1,j}, M_{i+1,j}, \ldots, M_{k,j}$. By chain rule of mutual information,

$$\mathbb{I}(M_{i,j}; \Pi(U_j) \mid (\Sigma, J)) = \mathbb{I}(M_{i,j}; \Pi(U_j) \mid \Sigma, J) + \mathbb{I}(M_{i,j}; \Pi(U_j) \mid M_{i,j}, \Sigma, J),$$

since $\mathbb{I}(M_{i,j}; \Pi(U_j) \mid M_{i,j}, \Sigma, J) = 0$, as $\Pi(U_j) \perp M_{i,j} \mid M_{i,j}, \Sigma, J$. 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 $M_{i,j} \mid \Sigma, J$ 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 $M_{i,j}$, they need to reveal enough information about every induced matching in $G_i$ in order to reveal enough information about the (unknown) matching $M_{i,j}$.

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

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

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

$$\mathbb{I}(M_{i,j}; \Pi(U_j) \mid (\Sigma, J)) \leq \mathbb{I}(M_{i,j}; \Pi(U_j) \mid \Sigma_i, J).$$

We bound the RHS of the above equation as follows,

$$\mathbb{I}(M_{i,j}; \Pi(U_j) \mid \Sigma_i, J) = \mathbb{I}(M_{i,j}; \Pi(U_j) \mid \Sigma_i, J = j) = \frac{1}{t} \sum_{j=1}^{t} \mathbb{I}(M_{i,j}; \Pi(U_j) \mid \Sigma_i),$$

where the second equality is as the distribution of $(M_{i,j}, \Pi(U_j), \Sigma_i)$ is independent of the event $J = j$ (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 $(M_{i,j+1}, \ldots, M_{i,t}) \perp M_{i,j} \mid \Sigma_i$ and by Proposition 2.3,

$$\frac{1}{t} \sum_{j=1}^{t} \mathbb{I}(M_{i,j}; \Pi(U_j) \mid \Sigma_i) \leq \frac{1}{t} \sum_{j=1}^{t} \mathbb{I}(M_{i,j}; \Pi(U_j) \mid M_{i,j+1}, \ldots, M_{i,t}),$$

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

$$\frac{1}{t} \cdot \mathbb{H}(M_{i,j+1}, \ldots, M_{i,t}; \Pi(U_j) \mid \Sigma_i) \leq \frac{1}{t} \cdot \mathbb{H}(\Pi(U_j)),$$

finalizing the proof.

**Proof of Theorem 1.** Let $\pi$ be any protocol (deterministic or randomized) for the maximal matching problem over the distribution $D_{MM}$. 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

$$k \cdot r/6 \leq \mathbb{I}(M_{1,j}, \ldots, M_{k,j}; \Pi \mid (\Sigma, J)) \leq \mathbb{H}(\Pi(P)) + \frac{1}{t} \cdot \sum_{i=1}^{k} \mathbb{E}(\Pi(U_i)) \leq |P| \cdot b + \frac{kN \cdot b}{t} \leq Nb + \frac{k}{t} \cdot Nb = 2Nb.$$

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

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

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 $D_{MM}$ 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 $j^*$ and $\sigma$ 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 \( \Omega \left( \frac{n^{1/2}}{c^{n^{1/2}}} \right) \) 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 \( \mathcal{D}_{MM} \)**

We design a reduction that given a graph \( G \sim \mathcal{D}_{MM} \), 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 \( \mathcal{D}_{MM} \), 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 \( \mathcal{M}_{i,j}^{RS} \) to be a matching on vertices of \( G_i \) that is a copy of the \( j \)-th induced matching of \( G^{RS} \) before dropping its edges in \( G_i \) randomly (hence, \( \mathcal{M}_{i,j}^{RS} \) is a superset of the induced matching of \( G_i \)). Also note that \( \mathcal{M}_{i,j}^{RS} \) for every \( i \in [k] \) is supported on unique vertices. By construction, \( \mathcal{M}_{i,j}^{RS} \) is only a function of \( \sigma \) and \( j \) (to determine which matching to pick, and which vertices in \( G \) are endpoints of this reduction).

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

**Reduction from maximal matching on \( \mathcal{D}_{MM} \):**

1. Suppose \( G \) is an \( n \)-vertex graph sampled from \( \mathcal{D}_{MM} \). The players collectively create the graph \( H \) on \( 2n \) vertices as follows:
   a. Each vertex \( u \in G \) creates two copies \( u^f \) and \( u^r \) of the same vertex, and connect \( u^f \) to \( v^f \) 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 \( V^f \) and \( V^r \).
   b. Each public vertex \( u \) in \( G \) adds an edge between \( u^f \) and \( v^f \), and also between \( u^r \) and \( v^r \), 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 \( u^f \) and \( u^r \) and sending their messages to the referee. The referee computes the maximal independent set \( M \) of \( H \).

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

4. If \( |M^f| \geq |M^r| \), the referee outputs the pre-image of edges of \( M^f \) in \( G \) as the final matching (that is, for every \( (u^f, v^f) \in M^f \), 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 \( P^f, P^r, \) and \( U^f, U^r \), 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 \( S \cap P^f = \emptyset \) (resp. \( S \cap P^r = \emptyset \)). Let \((u, v)\) be any edge in any \( \mathcal{M}_{i,j}^{RS} \) for \( i \in [k] \). Then \((u, v)\) survived the random sampling (in \( \mathcal{D}_{MM} \)) in \( G \) if and only if not both of \( u^f, v^f \) belong to \( S \) (resp. not both of \( u^r, v^r \) belong to \( S \)).**

**Proof.** We only prove the lemma for \( P^f \); the case for \( P^r \) follows by symmetry.

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

**Proof of Theorem 2.** Let \( \pi \) be any protocol (deterministic or randomized) for maximal independent set and let \( b \) denote the worst-case length of messages communicated by any player.
As explained in the reduction, each vertex \( u \in G \) can create the neighborhood of both \( u^e \) and \( u^f \) correctly in \( H \), and thus simulate \( \pi \) for them in \( H \) consistently with at most \( 2 \cdot b \) communication from \( u \). By definition, \( \pi \) outputs a correct \( M \) with probability at least 0.99. Whenever this happens, by construction of \( H \), we know that at least one of \( M \cap P^f \) or \( M \cap P^r \) should be empty (since all vertices in \( P^f \) and \( P^r \) are connected to each other). Conditioned on this event, by Lemma 4.1, at least one of \( M^e \) or \( M^f \) contains all edges between unique vertices \( U^e \) or \( U^f \), 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 \( 2b = \Omega(n^{1/2} / \epsilon^{2k} \sqrt{\log n}) \) 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


