These lecture notes have not undergone rigorous peer-review. Please email quanquan.liu@yale.edu if you see any errors.

1 Introduction

Today we'll discuss estimating the size of the maximum cardinality matching in an insertion-only stream. Our lecture will focus on the recent result of McGregor and Vorotnikova [MV18] which builds off the result of Cormode, Jowhari, Monemizadeh, and Muthukrishnan [CJMM16]. In *insertion-only* streams, edges arrive in arbitrary order and are inserted into the graph. Elements appear one at a time and if you do not store an element in your (small) memory when you see it, you cannot recall that element until the next pass of the stream. The goal is to minimize the number of passes of the stream and the space used for storing elements from the stream to compute the property of interest. Today's lecture focuses on streaming algorithms for maximum matching in low arboricity graphs. A matching in a graph is a set of edges where no two edges in the set share an endpoint. The maximum matching in the graph is a matching of maximum cardinality.

2 Properties of Arboricity

Recall from our previous lecture that the ***arboricity*** of a graph is defined as the minimum number of forests to decompose the edges of a graph. The arboricity of a graph is related to the *degeneracy* and *density* of the graph. Typically, the arboricity is denoted by α. The arboricity of the graph posesses a number of interesting properties:

- By the Nash-Williams theorem, we can show that that $\alpha = \max_{S \subseteq V} \left\{ \lceil \frac{E(S)}{|S|-1} \rceil \right\}$ where $E(S)$ is the set of edges in the induced subgraph given by S.
- The arboricity of planar graphs is ≤ 3.
- The arboricity of any subgraph of G is at most the arboricity of G.
- In a graph with α arboricity and n nodes, the number of edges in the graph is $\leq n\alpha$.

3 Streaming Maximum Matching in Low Arboricity Graphs

The general strategy for streaming approximation algorithms is the following:

1. Figure out the quantity we can approximate using sampling where the quantity also gives an estimate of the property you want to eventually compute.
2. Approximate the desired quantity via sampling and prove concentration bounds.

Let $M(G)$ be the maximum size of a matching in input graph G and $M(G)$ is the property we want to compute. For a stream of length m, let G^t be the prefix of the stream consisting of the first t elements. Let B_u^t be the *last* $\alpha + 1$ edges incident to $u \in V$ that appears in G^t. Let E_α^t be the set of edges $\{u, v\}$ in the stream where $\{u, v\} \in G^t$. That is, E_α^t consists of the set of edges $\{u, v\}$ where the number of edges incident to u and v that appear in the stream after $\{u, v\}$ is at most α. We say an edge $\{u, v\}$ is ***good*** if $\{u, v\} \in B_u \cap B_v$, and an edge is ***wasted*** if $\{u, v\} \in B_u \oplus B_v = (B_u \cup B_v) \setminus (B_u \cap B_v)$. Then, E_α^t is precisely the set of good edges in G^t. In other words, $E_\alpha^t = \bigcup_{u \neq v \in V} \left(B_u^t \cap B_v^t\right)$.

The size of E_α is the quantity we want to approximate using our algorithm. So, we first relate $|E_\alpha|$ to $M(G)$.

Lemma 3.1. $M(G^t) \leq |E^t_\alpha| \leq (\alpha + 2) \cdot M(G^t)$.

Proof. We'll first prove the right-hand side of this expression. To do this, we define a fractional matching using E^t_α. Let $Y_e = \frac{1}{\alpha+1}$ if $e \in E^t_\alpha$ and $Y_e = 0$ otherwise. Then, $\{Y_e\}_{e\in E}$ is a fractional matching with maximum weight $\frac{1}{\alpha+1}$.[1] We now show a corollary of Edmond's matching polytope theorem. Edmond's matching polytope theorem implies that if the weight of a fractional matching on any induced subgraph $S \subseteq G$ is at most $\frac{|S|-1}{2}$, then the weight on the entire graph is at most $M(G)$. Now, we show the following corollary:

Corollary 3.2. *Let $\{Y_e\}_{e\in E}$ be a fractional matching where the maximum weight on any edge is ε. Then, $\sum_{e\in E} Y_e \leq (1 + \varepsilon) \cdot M(G)$.*

Proof. Let S be an arbitrary subset of vertices, and let $E(S)$ be the edges in the induced subgraph of S. We know that $|E(S)| \leq \frac{|S|(|S|-1)}{2}$ and by the definition of fractional matching, $\sum_{e\in E(S)} Y_e = \frac{\sum_{u\in V}\sum_{v\in N(v)} f(\{u,v\})}{2} = \frac{\sum_{u\in V} 1}{2} = \frac{|S|}{2}$ where $\sum_{v\in N(v)} f(\{u, v\}) = 1$ by the constraints of the fractional matching. Thus, we know that

$$\begin{aligned}\sum_{e\in E(S)} &\leq \min\left(\frac{|S|}{2}, \frac{\varepsilon|S|(|S|-1)}{2}\right)\\ &\leq \frac{|S|-1}{2} \cdot \min\left(\frac{|S|}{|S|-1}, \varepsilon|S|\right)\\ &\leq \frac{|S|-1}{2} \cdot (1+\varepsilon).\end{aligned}$$

We can show the last inequality by considering two cases:

- If $\frac{|S|}{|S|-1} \leq \varepsilon|S|$, then

$$\begin{aligned}1 + \frac{1}{|S|-1} &\leq \varepsilon + 1\\ \frac{|S|}{|S|-1} &\leq 1 + \varepsilon.\end{aligned}$$

- If $\varepsilon|S| \leq \frac{|S|}{|S|-1}$, then

$$\begin{aligned}\varepsilon &\leq \frac{1}{|S|-1}\\ \varepsilon|S| - \varepsilon &\leq 1\\ \varepsilon|S| &\leq 1 + \varepsilon.\end{aligned}$$

Finally, let $Z_e = \frac{Y_e}{1+\varepsilon}$. We can use Edmond's polytope theorem to show that $\sum_{e\in E} Z_e \leq \frac{|S|-1}{2} \leq M(G)$ and so $\sum_{e\in E} Y_e \leq (1 + \varepsilon)\sum_{e\in E} Z_e \leq (1 + \varepsilon) \cdot M(G)$. We have proven our corollary. □

Using the above corollary with maximum weight $\frac{1}{\alpha+1}$ implies that $\sum_{e\in E} Y_e = \frac{|E^t_\alpha|}{\alpha+1} \leq (1+\frac{1}{\alpha+1})M(G) = \frac{\alpha+2}{\alpha+1} \cdot M(G)$. Hence, we have that $|E^t_\alpha| \leq (\alpha + 2) \cdot M(G)$.

Now, we prove the left inequality. To prove the left inequality, let H be the set of vertices in G^t with degree at least $\alpha + 1$. These are the ***heavy*** vertices. We also define the following variables:

[1] A fractional matching is defined to a set of weights $f(e)$ on every edge e where $\forall v \in V : \sum_{e\ni v} f(e) \leq 1$.

- $w :=$ the number of good edges with *no* endpoints in H,
- $x :=$ the number of good edges with *exactly one* endpoint in H,
- $y :=$ the number of good edges with *two* endpoints in H,
- $z :=$ the number of *wasted* edges with *no* endpoints in H.

First, $|E^t_\alpha| = w + x + y$. Now, we show the following additional equalities. We will first calculate the number of edges in the B_u of every $u \in H$ in terms of the variables we defined above. Since every vertex $u \in H$ is heavy, $|B_u| = \alpha + 1$. Hence, $\sum_{u \in H} |B_u| = (\alpha + 1)|H|$. Every edge in each of these B_u must either be a good edge or a wasted edge since they have at least one endpoint, namely u, which has at most α edges incident to it in the rest of G^t. For each edge counted in x, it has one endpoint in H so it is counted in exactly one B_u in H. Furthermore, for each edge counted in z, it is also counted in the B_u of exactly one of its two endpoints since it is wasted (i.e. the other endpoint doesn't have at most α edges that come after it). This leaves every good edge counted in y which is counted in exactly two B_u's. Hence, $\sum_{u \in H} |B_u| = x + 2y + z = (\alpha + 1)|H|$.

Now, we can also compute $z + y \leq \alpha|H|$ since $z + y$ is a subset of the total number of edges in the induced subgraph consisting of H. Since we know that the graph has arboricity α, we also know that (by the properties of arboricity) that the induced subgraph consisting of H has at most $\alpha|H|$ edges. Hence, we can sum our inequalities: $x + 2y + z = (\alpha + 1)|H|$ and $-z - y \geq -\alpha|H|$ to obtain $x + y \geq |H|$. Finally, let E_L be the set of edges with no endpoints in H. Every edge in E_L is good and $w = |E_L|$. Therefore, $|E^t_\alpha| = w + x + y = |H| + |E_L|$. We can show that $|H| + |E_L| \geq M(G)$ since we can partition the set of edges in the maximum matching to edges in E_L and edges incident to H. All of the edges in E_L can be in the matching and at most one edge incident to each vertex in H is in the matching. We have successfully proven our lower bound: $|E^t_\alpha| \geq M(G)$. □

Now we have the following algorithm for estimating $|E^t_\alpha|$ for every $t \leq m$. For every sampled edge, $e = \{u, v\}$, the algorithm also stores counters c^u_e nad c^v_e for the degrees of u and v in the rest of the stream. This requires an additional factor of $O(\log(\alpha))$ space. Thus, the algorithm maintains the invariant that each edge stored in the sample is a good edge with respect to the current G^t. The algorithm is given below in Algorithm 1.

Theorem 1 (Multiplicative Chernoff Bound)**.** *The Chernoff Bound is a probabilistic inequality that provides an upper bound on the tail distribution of sums of independent random variables. There are many variants of the bound; we present the common multiplicative version. Formally, it is expressed as:*

$$\mathsf{Pr}(|X - \mu| \geq \varepsilon \cdot \mu) \leq 2 \exp\left(-\frac{\varepsilon^2 \mu}{3}\right),$$

where X is a random variable representing the sum of independent random variables in $[0, 1]$, and μ is the expected value of X.

Now, we show that we get an $(1 + \varepsilon)$-approximation with high probability. First, we note that $E^*_t = \max_{t' \leq t}(|E^{t'}_\alpha|)$ follows $M(G^t) \leq E^*_t \leq (2+\alpha)M(G^t)$ since $E^*_t \geq |E^t_\alpha|$, $M(G^{t'}) \leq M(G^t)$, and Lemma 3.1.

We now show our main lemma for our algorithm.

Lemma 3.3. *Algorithm 1 returns a $(1 + \varepsilon)$-approximation of $|E^t_\alpha|$ for every $t \leq m$ with high probability.*

Algorithm 1: Sampling $|E_\alpha^t|$

```
Function Alg(α, ε, n)
    $S \leftarrow \emptyset$
    $p \leftarrow 1$
    estimate $\leftarrow 0$
    for each $e = \{u, v\}$ do
        With probability $p$ add $e$ to $S$ and initialize counters $c_e^u \leftarrow 0$ and $c_e^v \leftarrow 0$
        for each edge $e' \in S$ do
            if $e'$ shares endpoint $w$ with $e$ then
                increment $c_{e'}^w$
                if $c_{e'}^w > \alpha$ then
                    remove $e'$ and corresponding counters from $S$
                end
            end
        end
        if $|S| > 80\varepsilon^{-2}\log(n)$ then
            $p \leftarrow p/2$
            Remove each edge in $S$ and counters with probability $1/2$
        end
        estimate $\leftarrow \max(\text{estimate}, |S|/p)$
    end
    return estimate
```

Proof. First, let $\tau = \frac{40 \log n}{\varepsilon^2}$ and define level i (starting with $i = 2$) to be the level that contains $|E_\alpha^t|$ if $2^{i-1} \cdot \tau < E_t^* \leq 2^i \cdot \tau$. Level $i = 1$ is defined as $0 \leq E_t^* \leq 2 \cdot \tau$. Suppose in a perfect situation, edge e is sampled in level i with probability $p_i' = \frac{1}{2^i}$ for $i \geq 2$ and with probability $p_1' = 1$ for $i = 1$.

In that case, we can use the multiplicative Chernoff bound to bound the probability that our estimate concentrates. However, it is not the case that p_i is guaranteed to be $\frac{1}{2^i}$ since it is determined adaptively by Algorithm 1. Hence, we need a slightly more complicated analysis than simply using the multiplicative Chernoff bound with p_i'. We consider two cases. If $p_i \geq p_i'$, then we can lower bound the probability of success by what we would obtain using p_i' and we can use the multiplicative Chernoff bound in this case. Now, suppose that $p_i < p_i'$; then, we show that this case never occurs in Algorithm 1 with probability at least $1 - \frac{1}{n^3}$. (We can amplify this probability to any $1 - \frac{1}{\text{poly}(n)}$ by changing the constant 80.) We now formally present these arguments.

If $p_i \geq p_i'$, then we use the multiplicative Chernoff bound to bound our probability of success. Note that since we consider E_t^* instead of $|E_\alpha^t|$ for every t, we need only consider the time stamps t where $E_t^* = |E_\alpha^t|$. Below, we have $\mu = p_i \cdot |E_\alpha^t|$. Let S^t be the sample of edges for G_t.

$$\begin{aligned}
\mathsf{Pr}\left[|X-\mu| \geq \varepsilon \cdot \mu\right] &\leq 2\exp\left(-\frac{\varepsilon^2\mu}{3}\right)\\
\mathsf{Pr}\left[|S^t - p_i \cdot |E^t_\alpha|| \geq \varepsilon \cdot p_i \cdot |E^t_\alpha|\right] &\leq 2\exp\left(-\frac{\varepsilon^2|E^t_\alpha|p_i}{3}\right)\\
&\leq 2\exp\left(-\frac{\varepsilon^2|E^t_\alpha|p'_i}{3}\right) \quad \text{by our assumption that } p_i \geq p'_i\\
&\leq 2\exp\left(-\frac{20\log n}{3}\right) \quad \text{by our definition of } p'_i \text{ so that } p'_i \cdot E^*_t \geq \tau/2\\
&\leq 2\exp\left(-6\log n\right) = \frac{1}{n^6}.
\end{aligned}$$

We can adjust the constant 80 in Algorithm 1 to ensure high probability.

Now, we consider the case when $p_i < p'_i$. We show that with probability at least $1 - \frac{1}{n^3}$, this case *does not occur*. We prove this via induction on t. For $t = 1$, $p_i = p'_i$ trivially since both equal 1. Now, we assume our claim holds for t and show it holds for $t + 1$. Suppose for the sake of contradiction that with probability greater than $\frac{1}{n^3}$, $p_i < p'_i$. First, note that E^*_t cannot decrease so p'_i cannot decrease as t increases. Then, the only reason that $p_i < p'_i$ is if we sample more than $\frac{80\log n}{\varepsilon^2}$ elements of E^t_α at probability $p_i = p'_i$. Again, we only consider cases where $|E^t_\alpha| = E^*_t$. Then, we can use the Chernoff bound to show that

$$\mathsf{Pr}\left[|S^t - p'_i \cdot |E^t_\alpha|| \geq C \cdot p'_i \cdot |E^t_\alpha|\right] \leq 2\exp\left(-\frac{C^2 p'_i \cdot |E^t_\alpha|}{3}\right).$$

for some C which we compute below.

Let $\mu = p'_i \cdot |E^t_\alpha|$. Note that if $\mu \geq \frac{10\log n}{\varepsilon^2}$, then we can directly set $C = \varepsilon$ and obtain our desired contradiction. Thus, we show the case when $\mu < \frac{10\log n}{\varepsilon^2}$. Since we sampled more than $\frac{80\log n}{\varepsilon^2}$ elements, it must be the case that $S^t \geq \frac{80\log n}{\varepsilon^2}$ and $C \geq \frac{S^t-\mu}{\mu} = \frac{S^t}{\mu} - 1 \geq \frac{7S^t}{8\mu}$. Substituting this C into our bound gives

$$\begin{aligned}
\mathsf{Pr}\left[|S^t - \mu| \geq C \cdot \mu\right] &\leq 2\exp\left(-\frac{C^2\mu}{3}\right)\\
&\leq 2\exp\left(-\frac{(\frac{7S^t}{8\mu})^2\mu}{3}\right)\\
&= 2\exp\left(-\frac{49(S^t)^2}{192\mu}\right)\\
&\leq 2\exp\left(-\frac{49 \cdot 640\log n}{192\varepsilon^2}\right)\\
&\leq \frac{1}{n^{160}}.
\end{aligned}$$

We now union bound over all possible $t \leq n^2$ to obtain our probability of failure is at most $\frac{1}{n^{158}}$. This contradicts our assumption and it must be the case that $p_i \geq p'_i$ for $t + 1$.

□

Now, combining Lemma 3.1 with Lemma 3.3 gives us our final theorem.

Theorem 3.4. *With probability at least $1 - \frac{1}{n^3}$,*[2] *and using $O\left(\frac{\log(n)\log(\alpha)}{\varepsilon^2}\right)$ space, we can find a $(1+\varepsilon)(2+\alpha)$-approximation of the size of the maximum matching in an one-pass, arbitrary-order stream where α is the (given) arboricity of the graph.*

References

[CJMM16] Graham Cormode, Hossein Jowhari, Morteza Monemizadeh, and Shanmugavelayutham Muthukrishnan. The sparse awakens: Streaming algorithms for matching size estimation in sparse graphs. *arXiv preprint arXiv:1608.03118*, 2016.

[MV18] Andrew McGregor and Sofya Vorotnikova. A simple, space-efficient, streaming algorithm for matchings in low arboricity graphs. In *1st Symposium on Simplicity in Algorithms (SOSA 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[2] Can be amplified to high probability by changing the constant 80 in Algorithm 1