Maximum Matching in Turnstile Streams

ESA 2015

Christian Konrad

Reykjavik University

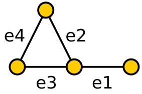
14.09.2015

Streaming Algorithms for Graph Problems

Insertion-only Streams (1999 -)

- Input stream: Sequence of edges of input graph $G = (V, E)$ with $n = |V|$ in arbitrary order

$$S = e_2 e_1 e_4 e_3$$



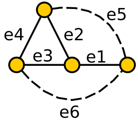
- Goal: Few passes (preferably one) algorithms with space $o(n^2)$
- Matchings, independent sets, cuts, graph sparsifiers, random walks, bipartiteness testing, counting triangles/subgraphs, ...

Dynamic/Turnstile Streams (2012 -)

- Input stream: Sequence of edge insertions/deletions, arbitrary order

$$S = e_4 e_3 e_5 \overline{e_5} e_2 e_6 \overline{e_2} e_2 e_1 \overline{e_6}$$

(arbitrary length)



- Goal: Few passes (preferably one) algorithm with space $o(n^2)$

Matching in Insertion-only Streams

Greedy Matching Algorithm

- Insert e into initially empty matching M if $M \cup \{e\}$ is a matching
- One-pass $\frac{1}{2}$-approximation streaming algorithm with space $\mathrm{O}(n \log n)$

Most Studied Graph Problem in the Streaming Model

Unweighted/weighted, one-pass/multi-pass, adversarial arrival order/random order

[Feigenbaum et al., Theo. Comp. Sci. 2005] [McGregor, APPROX 2005], [Ahn, Guha, ICALP 2011], [Eggert et al., Algorithmica 2012], [Goel et al., SODA 2012], [Kapralov, SODA 2013] [Zelke, Algorithmica 2012], [Epstein et al., STACS 2010], [Crouch, Stubbs, APPROX 2014], [Konrad et al., APPROX 2012], [Kapralov et al., SODA 2014], [Kapralov et al., SODA 2014], [Esfandiari et al., SODA 2015], . . .

Dynamic/Turnstile Streams

How well can we do if edge deletions are allowed?

(Open question from Bertinoro workshop 2014 on sub-linear algorithms)

Known Results on Dynamic Graph Streams

- **[Ahn, Guha, McGregor, SODA 2012]** Connectivity, bipartiteness, const. factor minimum weight spanning tree in $\mathrm{O}(n \log n)$ space

  $(1+\epsilon)$-approximate weighted matching with $\mathrm{O}(n^{1+1/p} \operatorname{poly} \epsilon^{-1})$ space and $\mathrm{O}(p \cdot \epsilon^{-2} \cdot \log \epsilon^{-1})$ passes
- **[Ahn, Guha, McGregor, APPROX 2013]**,**[Kapralov, Woodruff, PODC 2014]**, **[Kapralov et al., FOCS 2014]** Sparsifiers and spanners

Our Results

- **Upper Bound:** For every $0 \leq \epsilon \leq 1$: One-pass $\mathrm{O}(n^{\epsilon})$-approximation streaming algorithm with space $\tilde{\mathrm{O}}(n^{2-2\epsilon})$
- **Lower Bound:** For every $0 \leq \epsilon \leq 1$: Every one-pass $\mathrm{O}(n^{\epsilon})$-approximation streaming algorithm requires space $\Omega(n^{3/2-4\epsilon})$

Upper Bound

Main Algorithmic Technique: Linear Sketches

Linear Sketches and Turnstile Streams

- Turnstile stream: Updates to characteristic vector of edges
  - x vector of integers of size $\binom{|V|}{2}$, initially $x = (0, \dots, 0)$
  - Edge insertion e_i: $x \leftarrow x + (0, \dots, 0, 1, 0 \dots, 0)$ (ith unit vector)
  - Edge deletion e_i: $x \leftarrow x - (0, \dots, 0, 1, 0 \dots, 0)$
- Sketching algorithm:
  1. Choose sampling matrix A from dist. of matrices (randomized)
  2. Compute sketch $y = A \cdot x$ (deterministic) while processing the stream
  3. Post-processing: Compute output from sketch y

Linear Sketches are Universal

- All known turnstile algorithms are linear sketches
- **[Li, Nguyen, Woodruff, STOC 2014]** For every turnstile alg., there is one that *behaves* the same and can be implemented as linear sketch
- Size of sketch at most log-factor larger than space of original alg.

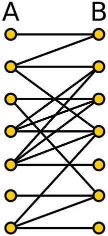
Upper Bound

L_0-**sampling**

- Given vector x, sample from non-zero coordinates u.a.r.
- **[Jowhari, Sağlam, Tardos, PODS 2011]** There is a turnstile algorithm with space $\mathrm{O}(\log^2(n) \log \frac{1}{\delta})$ that performs L_0-sampling with δ error

Algorithm (suppose perfect matching present)

- Input: Bipartite $G = (A, B, E)$ with $|A| = |B| = n$
- Let $A' \subseteq A$ u.a.r. subset of nodes of size $n^{1-\epsilon}$
- For every $a \in A'$: Sample $C \cdot n^{1-\epsilon} \log n$ times from set of incident edges
- Output largest matching M induced by sampled edges



Lemma
$\forall a \in A'$: at least $\min\{\deg_G(a), n^{1-\epsilon}\}$ incident edges sampled w.h.p.

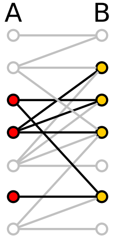
Upper Bound

L_0-sampling

- Given vector x, sample from non-zero coordinates u.a.r.
- **[Jowhari, Sağlam, Tardos, PODS 2011]** There is a turnstile algorithm with space $\mathrm{O}(\log^2(n) \log \frac{1}{\delta})$ that performs L_0-sampling with δ error

Algorithm (suppose perfect matching present)

- Input: Bipartite $G = (A, B, E)$ with $|A| = |B| = n$
- Let $A' \subseteq A$ u.a.r. subset of nodes of size $n^{1-\epsilon}$
- For every $a \in A'$: Sample $C \cdot n^{1-\epsilon} \log n$ times from set of incident edges
- Output largest matching M induced by sampled edges



$G[A' \cup B]$

Lemma

$\forall a \in A'$: at least $\min\{\deg_G(a), n^{1-\epsilon}\}$ incident edges sampled w.h.p.

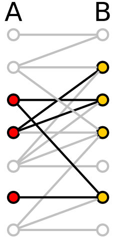
Upper Bound

L_0-**sampling**

- Given vector x, sample from non-zero coordinates u.a.r.
- **[Jowhari, Sağlam, Tardos, PODS 2011]** There is a turnstile algorithm with space $\mathrm{O}(\log^2(n) \log \frac{1}{\delta})$ that performs L_0-sampling with δ error

Algorithm (suppose perfect matching present)

- Input: Bipartite $G = (A, B, E)$ with $|A| = |B| = n$
- Let $A' \subseteq A$ u.a.r. subset of nodes of size $n^{1-\epsilon}$
- For every $a \in A'$: Sample $C \cdot n^{1-\epsilon} \log n$ times from set of incident edges
- Output largest matching M induced by sampled edges



sampled edges

Lemma
$\forall a \in A'$: at least $\min\{\deg_G(a), n^{1-\epsilon}\}$ incident edges sampled w.h.p.

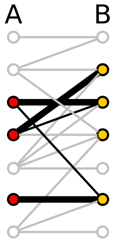
Upper Bound

L_0-sampling

- Given vector x, sample from non-zero coordinates u.a.r.
- **[Jowhari, Sağlam, Tardos, PODS 2011]** There is a turnstile algorithm with space $\mathrm{O}(\log^2(n) \log \frac{1}{\delta})$ that performs L_0-sampling with δ error

Algorithm (suppose perfect matching present)

- Input: Bipartite $G = (A, B, E)$ with $|A| = |B| = n$
- Let $A' \subseteq A$ u.a.r. subset of nodes of size $n^{1-\epsilon}$
- For every $a \in A'$: Sample $C \cdot n^{1-\epsilon} \log n$ times from set of incident edges
- Output largest matching M induced by sampled edges



matching M

Lemma

$\forall a \in A'$: at least $\min\{\deg_G(a), n^{1-\epsilon}\}$ incident edges sampled w.h.p.

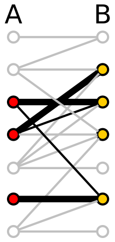
Upper Bound

L_0-sampling

- Given vector x, sample from non-zero coordinates u.a.r.
- **[Jowhari, Sağlam, Tardos, PODS 2011]** There is a turnstile algorithm with space $\mathrm{O}(\log^2(n) \log \frac{1}{\delta})$ that performs L_0-sampling with δ error

Algorithm (suppose perfect matching present)

- Input: Bipartite $G = (A, B, E)$ with $|A| = |B| = n$
- Let $A' \subseteq A$ u.a.r. subset of nodes of size $n^{1-\epsilon}$
- For every $a \in A'$: Sample $C \cdot n^{1-\epsilon} \log n$ times from set of incident edges
- Output largest matching M induced by sampled edges



Space Requirements: $\tilde{\mathrm{O}}(n^{2-2\epsilon})$

Lemma

$\forall a \in A'$: at least $\min\{\deg_G(a), n^{1-\epsilon}\}$ incident edges sampled w.h.p.

Upper Bound (2)

Let M^* be a maximum matching in G, then

$$\mathbb{E}\,|M^* \cap G[A' \cup B]| = |M^*|/n^{\epsilon}$$

Lemma We find a 1/2-approximation in graph $G[A' \cup B]$

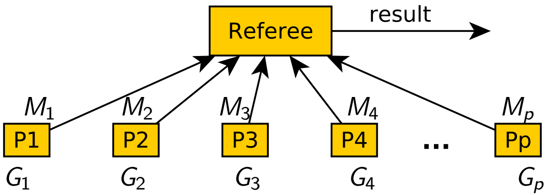
Proof Idea

1. Nodes $A_1 \subseteq A$ of degree at most $n^{1-\epsilon}$:
   - Graph $G[A_1 \cup B]$ entirely sampled
   - Let M_1 be a maximum matching in $G[A_1 \cup B]$
2. Nodes $A_2 = A \setminus A_1$ of degree at least $n^{1-\epsilon}$:
   - Hall's theorem: All A_2 vertices can be matched in graph $G[A_2 \cup B]$
3. $\mathbb{E}\max\{M_1, M_2\} \geq \mathbb{E}\,|M^* \cap G[A' \cup B]|/2$ □

Lower Bound

LBs via Simultaneous Communication Complexity

Simultaneous Communication Protocols



- Every party Pi holds a subgraph $G_i = (V, E_i)$ and $E_i \subseteq E$
- Every party sends message M_i to referee
- Referee computes output as a function of the messages

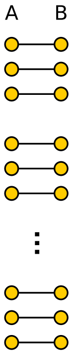
Turnstile Algorithm $\rightarrow$ Linear sketch $\rightarrow$ Sim. Communication protocol
(space s) (sketch size $\tilde{O}(s)$) (longest message $\tilde{O}(s)$)

Lemma

LB on size of any message M_i is LB on space of any turnstile algorithm

Hard Instance Construction

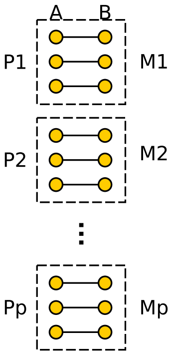
Global View



Perfect matching M in
bipartite graph $G = (A, B, E)$

Hard Instance Construction

Global View

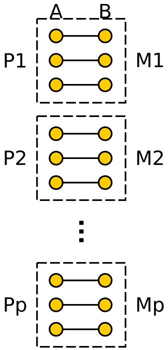


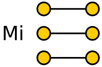
Each party should equally contribute to large global matching

$M = M_1 \cup M_2 \cup \cdots \cup M_p$

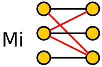
Hard Instance Construction

Global View





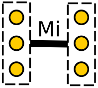
induced matching



alternative edges

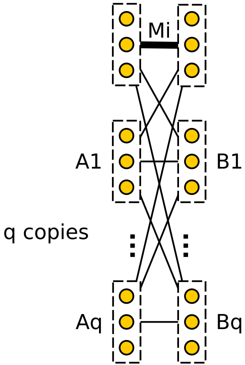
Matching M_i should be induced:
No alternative edges between
$A(M_i)$ and $B(M_i)$

Local View



Hard Instance Construction

Local View



Vertex groups interconnected with perfect matchings

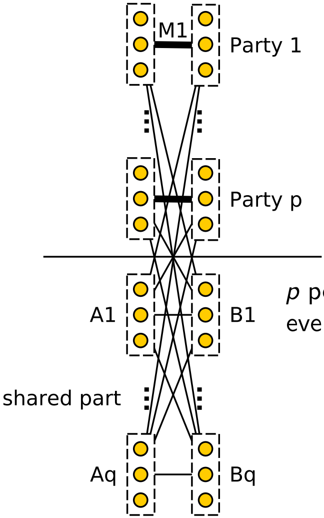
All perfect matchings are induced

By symmetry, M_i cannot be identified

Ruzsa-Szemerédi Graphs: Edge set can be partitioned into large induced matchings

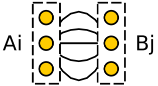
Hard Instance Construction

Global View



$p \gg q$

p perfect matchings between every pair of vertex groups



Summary: Lower Bound

Theorem For every $0 \leq \epsilon \leq 1$, every one-pass $\mathrm{O}(n^{\epsilon})$-approximation streaming algorithm requires space $\Omega(n^{3/2-4\epsilon})$.

$\rightarrow$ First lower bound for graph problems in the turnstile model

Summary and Conclusion

Our Result

In order to compute a n^ϵ-approximation to the maximum matching problem in the one-pass turnstile streaming model,

- space $\Omega(n^{3/2-4\epsilon})$ is required, and
- space $\tilde{O}(n^{2-2\epsilon})$ is sufficient.

Summary and Conclusion

Our Result

In order to compute a n^{ϵ}-approximation to the maximum matching problem in the one-pass turnstile streaming model,

- space $\Omega(n^{3/2-4\epsilon})$ is required, and
- space $\tilde{O}(n^{2-2\epsilon})$ is sufficient.

Independent Results

- **[Assadi et al. arXiv 2015]** Right answer: $n^{2-3\epsilon}$ (UB and LB)
- **[Chitnis et al. arXiv 2015]** Upper Bound: $n^{2-3\epsilon}$

Summary and Conclusion

Our Result

In order to compute a n^{ϵ}-approximation to the maximum matching problem in the one-pass turnstile streaming model,

- space $\Omega(n^{3/2-4\epsilon})$ is required, and
- space $\tilde{\mathrm{O}}(n^{2-2\epsilon})$ is sufficient.

Independent Results

- **[Assadi et al. arXiv 2015]** Right answer: $n^{2-3\epsilon}$ (UB and LB)
- **[Chitnis et al. arXiv 2015]** Upper Bound: $n^{2-3\epsilon}$

Open Problems

- Particular graph classes?
- Matching size?

Thank you.