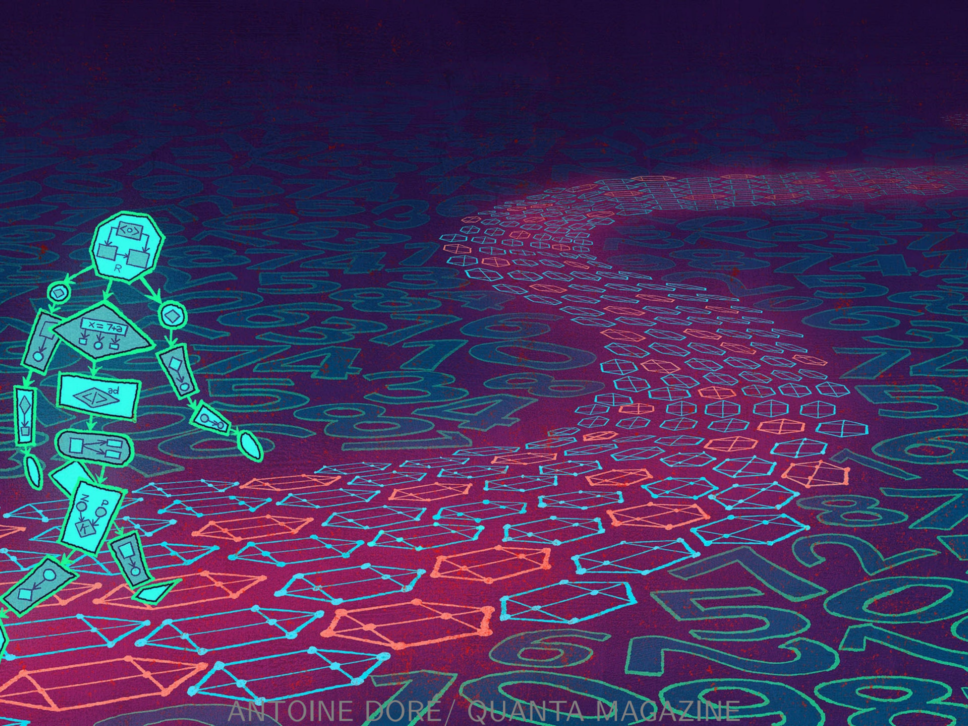
Topics in TCS

An introduction to data streaming

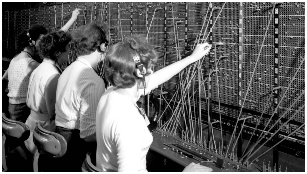
Raphaël Clifford


ANTOINE DORE / QUANTA MAGAZINE

Data streaming

This unit is about algorithms for processing data streams.

We will develop fast, small space data, typically but not always randomised data structures and algorithms.



For a small subset of the many applications, see e.g. Google's page on the Count-Min sketch[1].

[1]Some of the links are broken unfortunately but the application links work

What is in this half of the unit?

Subject:	Topics	Reference
What is streaming?	Introduction	
Probability overview	Markov, Chebyshev, Chernoff	MIT notes
Finding frequent elements	The Misa-Gries algorithm	Ch. 1
Counting distinct elements	The Tidemark algorithm	Ch. 2
Counting part II	The BJKST algorithm	Ch. 3
Approximate counting	The Morris counter	Ch. 4
Finding frequent items	CountSketch/Min Sketch	Ch. 5
Sparse recovery	Fingerprinting and hashing	Ch. 9
ℓ_0-sampling	Sample by frequency	Section 10.2

The set text is the Data Stream Algorithms by Chakrabati. A version without the word DRAFT is linked from the unit web page.

What is in the second half of the unit?

Subject:	Topics	Reference
Graph streams?	Connectivity, Bipartiteness	Ch. 14.2, 14.3
Shortest distances	Computing spanners	Ch. 14.4
Matchings I	Unweighted and weighted	Ch. 15
Matchings II	Multiple passes	Sec. 3 in [1]
Matchings III	Insertion-deletion streams	Sec. 5 in [2]
The AGM sketch	Connectivity with deletions	Ch. 16
Lower Bounds	Communication complexity	Ch. 18
Lower Bounds II	Yao's Lemma, INDEX problem	Ch. 18
Lower Bounds III	Further reductions	Ch. 18

[1] S. Kale, S. Tirodkar: Maximum Matching in Two, Three, and a Few More Passes Over Graph Streams. APPROX-RANDOM 2017.
`https://arxiv.org/pdf/1702.02559.pdf`

[2] C. Konrad: Maximum Matching in Turnstile Streams. ESA 2015.
`https://arxiv.org/pdf/1505.01460.pdf`

What is data streaming?

IP	Frequency
37.56.181.226	5
241.79.159.27	1
163.0.199.170	13
62.26.98.238	0
47.127.134.141	4
4.232.47.134	3
16.13.141.93	7

What is data streaming?

How many distinct IPs? What is the most frequent IP?
Estimate frequency of an IP? Randomly sample an IP. Etc.

IP	Frequency
37.56.181.226	5
241.79.159.27	1
163.0.199.170	13
62.26.98.238	0
47.127.134.141	4
4.232.47.134	3
16.13.141.93	7

What is data streaming?

How many distinct IPs? What is the most frequent IP?
Estimate frequency of an IP? Randomly sample an IP. Etc.

IP	Frequency
37.56.181.226	5
241.79.159.27	1
163.0.199.170	13
62.26.98.238	0
47.127.134.141	4
4.232.47.134	3
16.13.141.93	7

small space

What is data streaming?

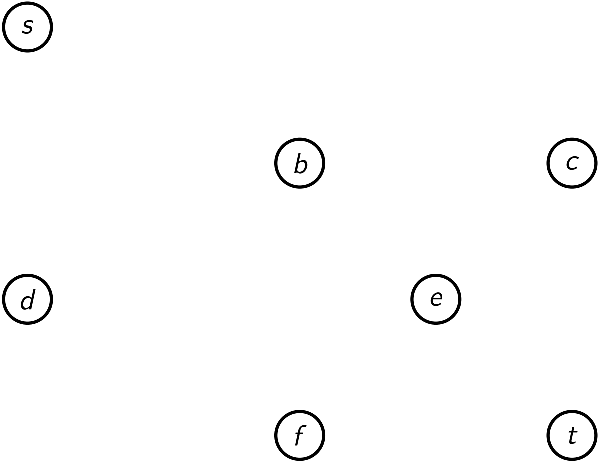
How many distinct IPs? What is the most frequent IP?
Estimate frequency of an IP? Randomly sample an IP. Etc.

IP	Frequency
37.56.181.226	5
241.79.159.27	1
163.0.199.170	13
62.26.98.238	0
47.127.134.141	4
4.232.47.134	3
16.13.141.93	7

small space

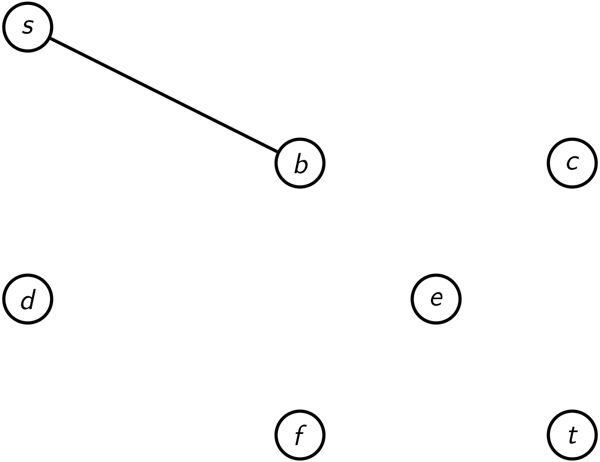
one-pass

Streaming graphs



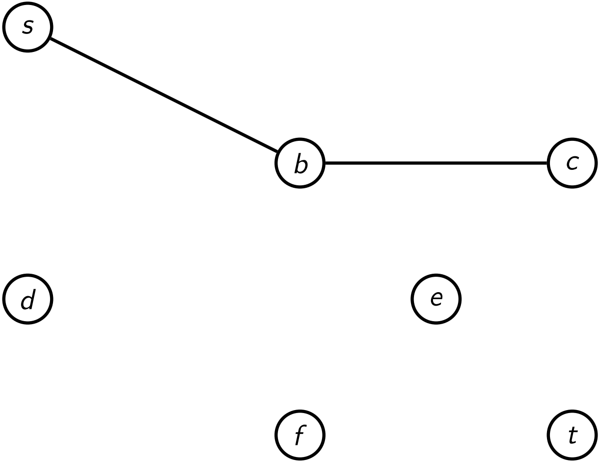
Streaming graphs

Add edge



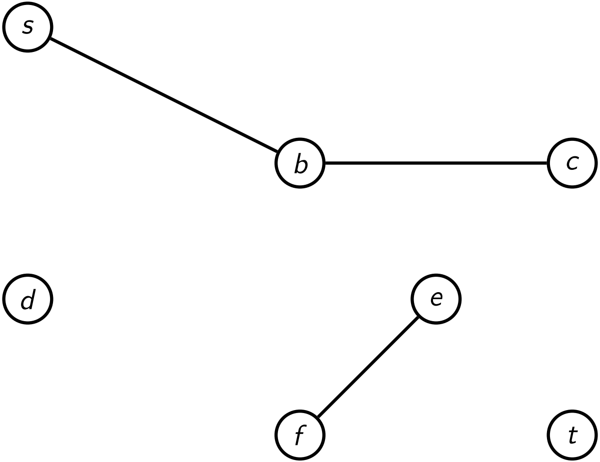
Streaming graphs

Add edge



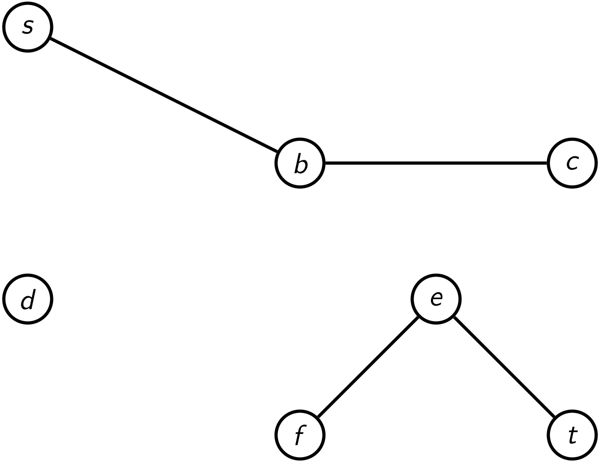
Streaming graphs

Add edge



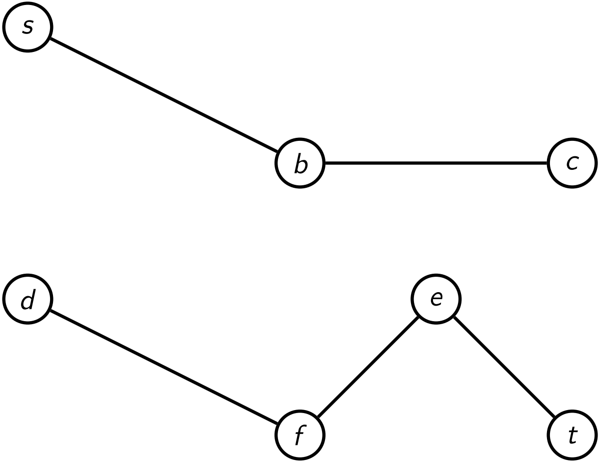
Streaming graphs

Add edge



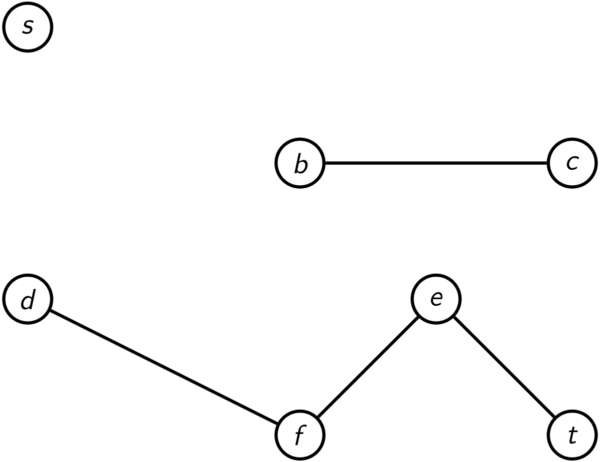
Streaming graphs

Add edge



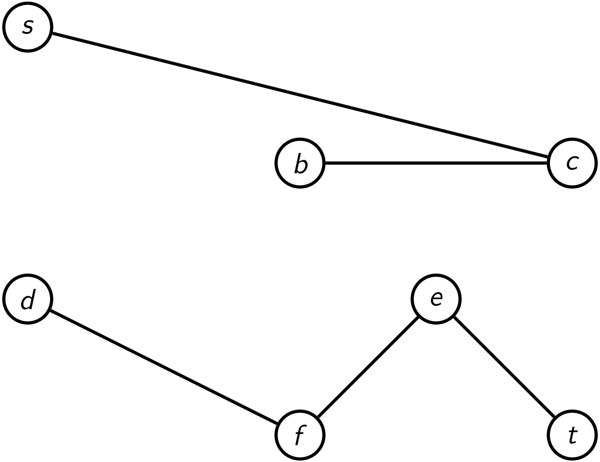
Streaming graphs

Delete edge



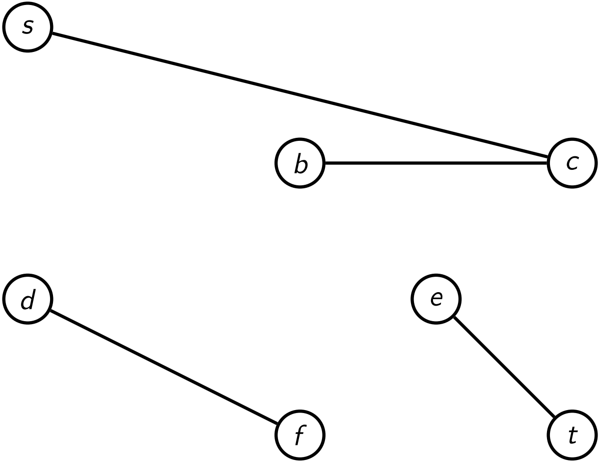
Streaming graphs

Add edge



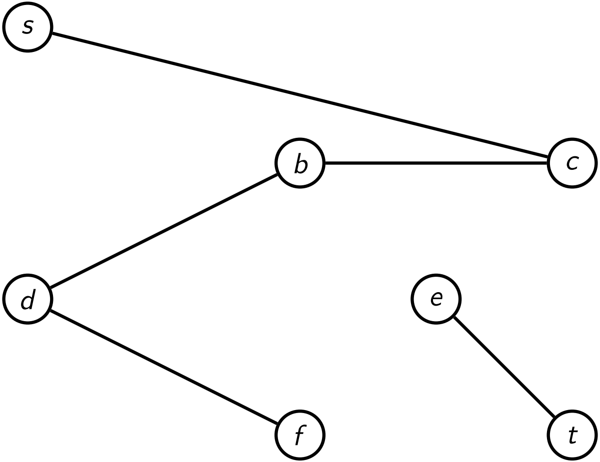
Streaming graphs

Delete edge



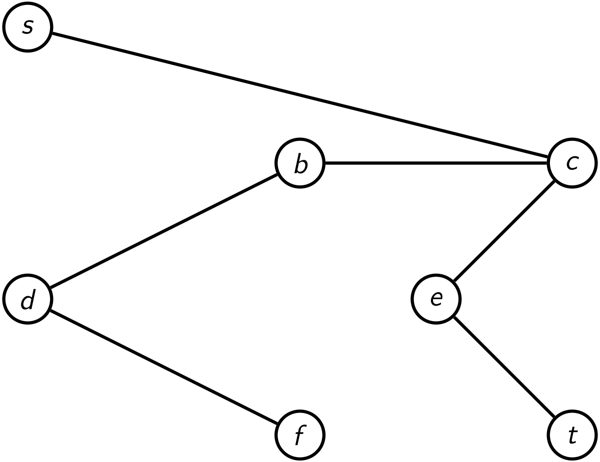
Streaming graphs

Add edge



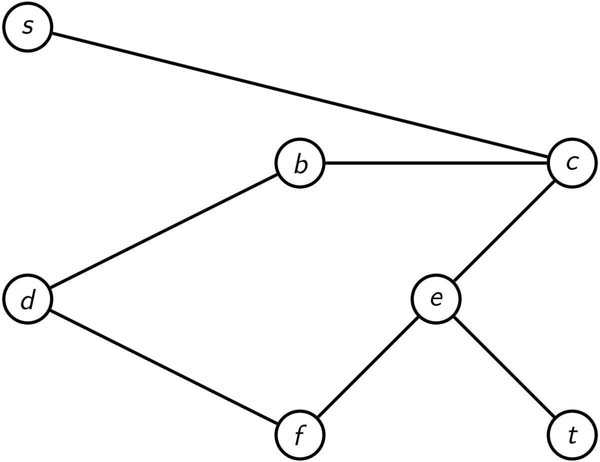
Streaming graphs

Add edge. s is connected to t.



Streaming graphs

Add edge. s is connected to t.



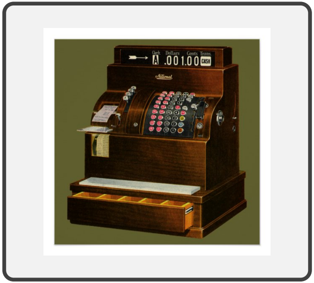
The cash register and turnstile models

(a) cash register

(b) turnstile

- Streaming elements may have an associated count. For example, two apples or eleven copies of IP address 37.56.181.226.

The cash register and turnstile models

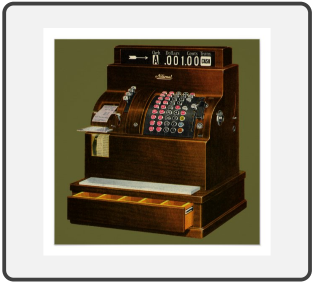


(a) cash register

(b) turnstile

- Streaming elements may have an associated count. For example, two apples or eleven copies of IP address 37.56.181.226.
- In the cash register streaming model counts are always non-negative.

The cash register and turnstile models



(a) cash register

(b) turnstile

- Streaming elements may have an associated count. For example, two apples or eleven copies of IP address 37.56.181.226.
- In the cash register streaming model counts are always non-negative.
- In the turnstile streaming the count may be negative or positive. For example we may remove copies of an IP address as well as adding copies or in a graph we may remove edges as well as add them.

Advantages of streaming algorithms

- In an internet router, for example, we never be able to store all the data and may want answers to be produced quickly.

Advantages of streaming algorithms

- In an internet router, for example, we never be able to store all the data and may want answers to be produced quickly.
- These properties may be desirable:

Advantages of streaming algorithms

- In an internet router, for example, we never be able to store all the data and may want answers to be produced quickly.
- These properties may be desirable:

  1. One-pass. This means we never go back and look at data in the past.

Advantages of streaming algorithms

- In an internet router, for example, we never be able to store all the data and may want answers to be produced quickly.
- These properties may be desirable:

  1. One-pass. This means we never go back and look at data in the past.
  2. Small space. This means we use much less space than it takes to store the whole input.

Advantages of streaming algorithms

- In an internet router, for example, we never be able to store all the data and may want answers to be produced quickly.
- These properties may be desirable:
    1. One-pass. This means we never go back and look at data in the past.
    2. Small space. This means we use much less space than it takes to store the whole input.
    3. Near linear time. This means near constant time per arriving token.

Advantages of streaming algorithms

- In an internet router, for example, we never be able to store all the data and may want answers to be produced quickly.
- These properties may be desirable:

    1. One-pass. This means we never go back and look at data in the past.
    2. Small space. This means we use much less space than it takes to store the whole input.
    3. Near linear time. This means near constant time per arriving token.

- If the data set is massive, fast, small space, one-pass algorithms may be needed even if it is not being streamed.

Is what we want to do possible?

- Sometimes there are proofs that an exact guaranteed correct answer cannot be given in sublinear space. This is true even more so for one-pass algorithms.

Is what we want to do possible?

- Sometimes there are proofs that an exact guaranteed correct answer cannot be given in sublinear space. This is true even more so for one-pass algorithms.

- Some lower bound proofs will be shown at the end of the unit.

Is what we want to do possible?

- Sometimes there are proofs that an exact guaranteed correct answer cannot be given in sublinear space. This is true even more so for one-pass algorithms.
- Some lower bound proofs will be shown at the end of the unit.
- Where exact and provably correct answers can't be given we will instead show approximate and/or randomised solutions which are correct with high probability.

Is what we want to do possible?

- Sometimes there are proofs that an exact guaranteed correct answer cannot be given in sublinear space. This is true even more so for one-pass algorithms.
- Some lower bound proofs will be shown at the end of the unit.
- Where exact and provably correct answers can't be given we will instead show approximate and/or randomised solutions which are correct with high probability.
- For example, answers that with 90% probability are within 10% of the correct value.

What do I need to do well this unit?

- The unit will use discrete probability to bound the probability of error of the various algorithms.

What do I need to do well this unit?

- The unit will use discrete probability to bound the probability of error of the various algorithms.

- Please reread the probability slides from Advanced Algorithms (linked from the unit web page). A good understanding of these will be expected.

What do I need to do well this unit?

- The unit will use discrete probability to bound the probability of error of the various algorithms.

- Please reread the probability slides from Advanced Algorithms (linked from the unit web page). A good understanding of these will be expected.

- All the probability needed is also covered in chapters 14-19 of (probability notes). You will not need all of this (in particular you won't need to learn any probability distributions except for the uniform distribution) but you should read it and keep it to hand.

What do I need to do well this unit?

- The unit will use discrete probability to bound the probability of error of the various algorithms.

- Please reread the probability slides from Advanced Algorithms (linked from the unit web page). A good understanding of these will be expected.

- All the probability needed is also covered in chapters 14-19 of (probability notes). You will not need all of this (in particular you won't need to learn any probability distributions except for the uniform distribution) but you should read it and keep it to hand.

- The unit set text is by Chakrabati and the latest version can be found at (here). A version without the word DRAFT is on the unit web page.

Other readings and related courses

- Andrew McGregor's 2012 course from the University of Massachusetts, Amherst (McGregor).
- Alexandr Andoni's 2015 course from the University of Columbia (Andoni).
- Indyk and Nelson's 2017 course from Harvard University (Harvard).