CMSC5741 Big Data Tech. & Apps.

Lecture a Streams

Prof. Michael R. Lyu

Computer Science & Engineering Dept.

The Chinese University of Hong Kong

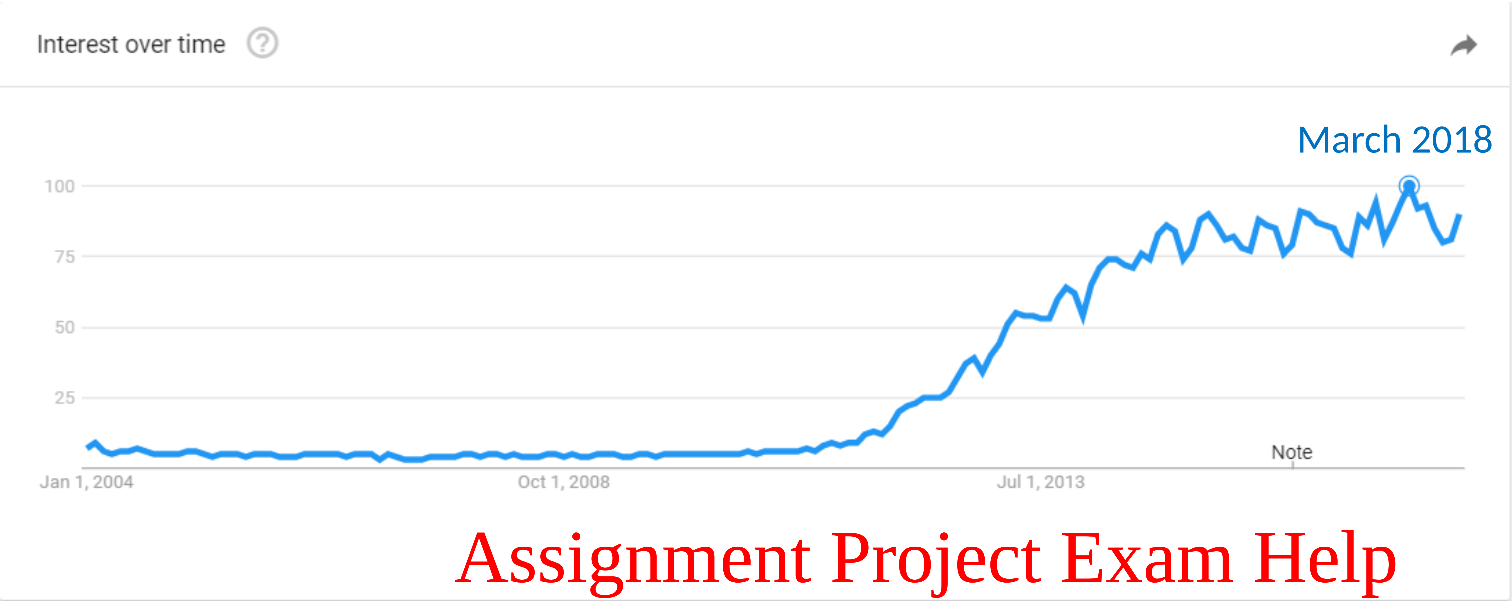
Motivation

- In many data mining situations, we know the entire data set in advance

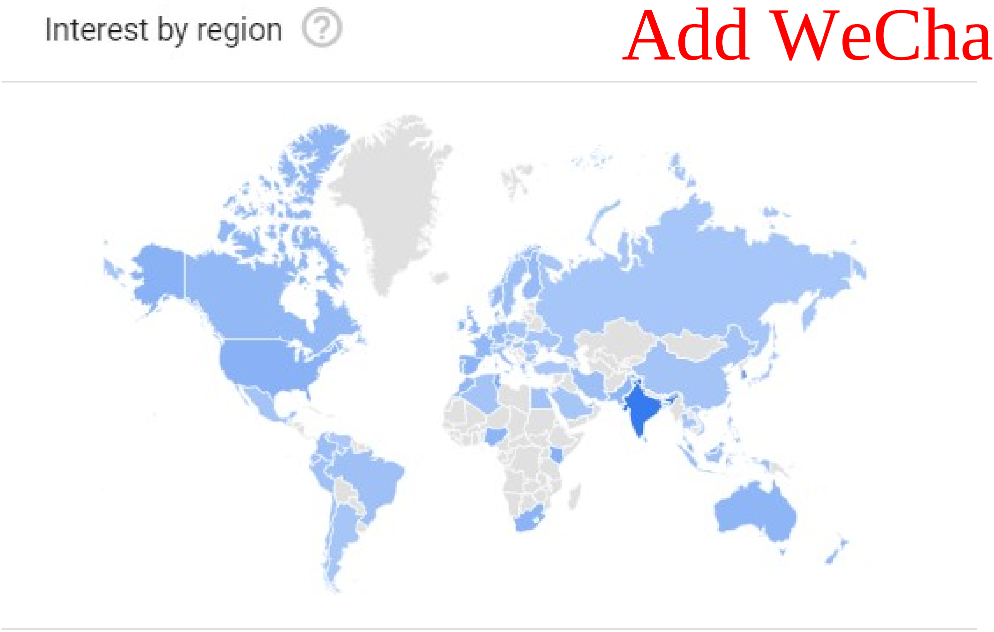
- Stream Man nt when the input rate is controlled e :
  - Google queries
  - Twitter and Facebook status updates
- We can think of the data as infinite and non-stationary (the distribution changes over time)

Google Trends

When we search for "big data"



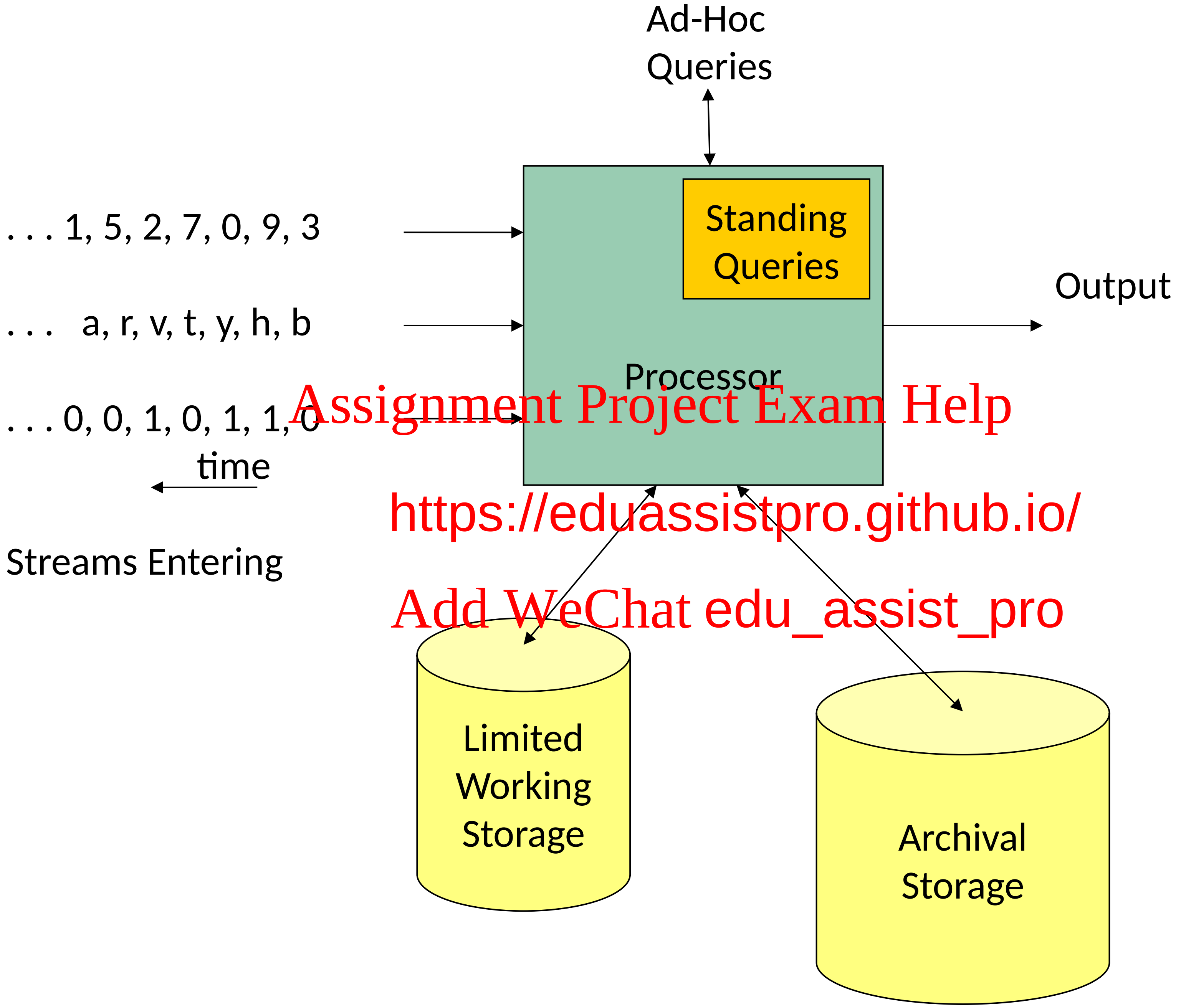
March 2018



Election 2016: Trump vs Clinton

The Stream Model

- Input tuples (e.g., [user, query, time]) enter at a rapid rate, at one or more input ports
- The system c ntire stream accessibly
- How do you make critic tions about the stream using a limited amount of (secondary) memory?


Ad-Hoc
Queries
Standing
Queries
Output
. . . 1, 5, 2, 7, 0, 9, 3
. . .   a, r, v, t, y, h, b
. . . 0, 0, 1, 0, 1, 1, 0
time
Processor
Streams Entering
Limited
Working
Storage
Archival
Storage

Problems on Data Streams

- Types of queries one wants on answer on a stream:

  - Sampling da
    - Construct
  - Queries over sliding win
    - Number of items of type *x* in the last *k* elements of the stream
  - Filtering a data stream
    - Select elements with property *x* from the stream

Problems on Data Streams

- Types of queries one wants on answer on a stream:
  - Counting di
    - Number of last *k* elements of the stream
  - Estimating moments
    - Estimate avg./std. dev. of last *k* elements
  - Finding frequent elements

Applications (1)

- Mining query streams
  - Google wants to know what queries are more frequent to

- Mining click
  - Yahoo! wants to know s pages are getting an unusual number of hits in the past hour

- Mining social network news feeds
  - E.g., look for trending topics on Twitter, Facebook

Applications (2)

- Sensors Networks
  - Many sensors feeding into a central controller
- Telephone c
  - Data feeds into custome ell as settlements between tel ompanies
- IP packets can be monitored at a switch
  - Gather information for optimal routing
  - Detect denial-of-service attacks

Outline

- Sampling from a Data Stream
- Queries over a (long) Sliding Windows
- Filtering Dat
- Counting Distinct Eleme
- Computing Moments
- Counting Itemsets

Outline

- Sampling from a Data Stream
- Queries over a (long) Sliding Windows
- Filtering Dat
- Counting Distinct Eleme
- Computing Moments
- Counting Itemsets

Sampling from a Data Stream

- Since we cannot store the entire stream, one obvious approach is to store a sample
- Two different
  - Sample a fixe ts in the stream (say 1 in 10)
  - Maintain a random sample of fixed size over a potentially in finite stream
    - At any "time" *t* we would like a random sample of *n* elements. For all *t*, each of *n* elements seen so far has equal prob. of being sampled

Sampling a Fixed Proportion

- Problem 1: Sampling fixed proportion
- Scenario: Search engine query stream
  - Stream of tup
  - Answer ques n did a user run the same query on two differen
  - Have space to store 1/10th of query stream
- Naive solution:
  - Generate a random integer in [0..9] for each query
  - Store the query if the integer is 0, otherwise discard

Problem with Naive Approach

- Simple question: What fraction of queries by an average user are duplicates?
- Suppose each user issues s queries once and d queries twice (total of *s* ate is p
  - Correct answer:
  - Sample will contain *sp* of the si ries and *2dp* of the duplicate queries at least once
  - But only dp^2 pairs of duplicates
    - $dp^2 = p * p * d$
  - Of d "duplicates" *2p(1-p)d* appear once
    - $2p(1-p)d = ((p*(1-p))+((1-p)*p))*d$
  - So the sample-based answer is: $dp^2/(sp+dp^2+ 2p(1-p)d)$

Problem with Naive Approach

- A concrete example:
  - Query stream: 1, 2, 3, 4, 5, 6, 7, 7, 8, 8
  - Sample 50% s case
  - Correct ans re duplicates
  - If our sample is 1, 2, 3, 4 % are duplicates
  - If our sample is 6, 7, 7, 8, 8, then 67% are duplicates
  - What is the expectation of fraction of duplicates if we use sample-based method?

Answer: 1/9 Solution?

Solution: Sample Users

- Pick 1/10th of users and take all their searches in the sample
- Use a hash f d the user name or use 10 buckets

- Generalized: Pick 1/dth of users, we need to use d buckets

Generalized Solution

- Stream of tuples with keys:
  - Key is some subset of each tuple's components
    - E.g., tuple i ey is user
  - Choice of k ation
- To get a sample of size *a* :
  - Hash each tuple's key uniformly into *b* buckets
  - Pick the tuple if its hash value is at most *a*

    (h(x) = 1, 2, …, a)

Maintaining a Fixed-size Sample

- Problem 2: Fixed-size sample
- Suppose we need to maintain a sample *S* of size exactly *s* (*s* is fi out of *S*=100 space)
  - E.g., main memory size const

- Why? Don’t know length of stream in advance
  - In fact, stream could be infinite
- Suppose at time *t* we have seen *n* items
  - Ensure each item is in the sample *S* with equal prob. *s*/*n*

Solution: Fixed Size Sample

- Algorithm:
  - Store all the first s elements of the stream to S

  - Suppose we have seen *n* elements, and now the *n+1*th element arrive

    - With prob. *s/n+1*, pick the *n+1* lse discard it

    - If we picked the *n+1*th element aces one of the *s* elements in the sample *S*, picked uniformly at random
- Claim: This algorithm maintains a sample *S* with the desired property, i.e., each item is in the sample *S* with equal prob.

Proof: By Induction

- We prove this by induction:
  - Assume that after *n* elements, the sample contains each element seen so far with prob. *s/n*

  - We need to ng element *n+1* the sample maintains th

    - Sample contains each element seen so far with prob. *s/(n+1)*
  - Obviously, after we see *n=s* elements the sample has the wanted property
    - Each out of n=s elements is in the sample with prob. *s/s=1*

Proof: By Induction

- After n elements, the sample S contains each element seen so far with probability s/n
- Now element
- For elements bility of remaining in S is:

- At time n tuples in S were there with prob. s/n
- Time n→n+1 tuple stayed in S with prob. n/(n+1)
- So prob. tuple is in S at time n+1 =

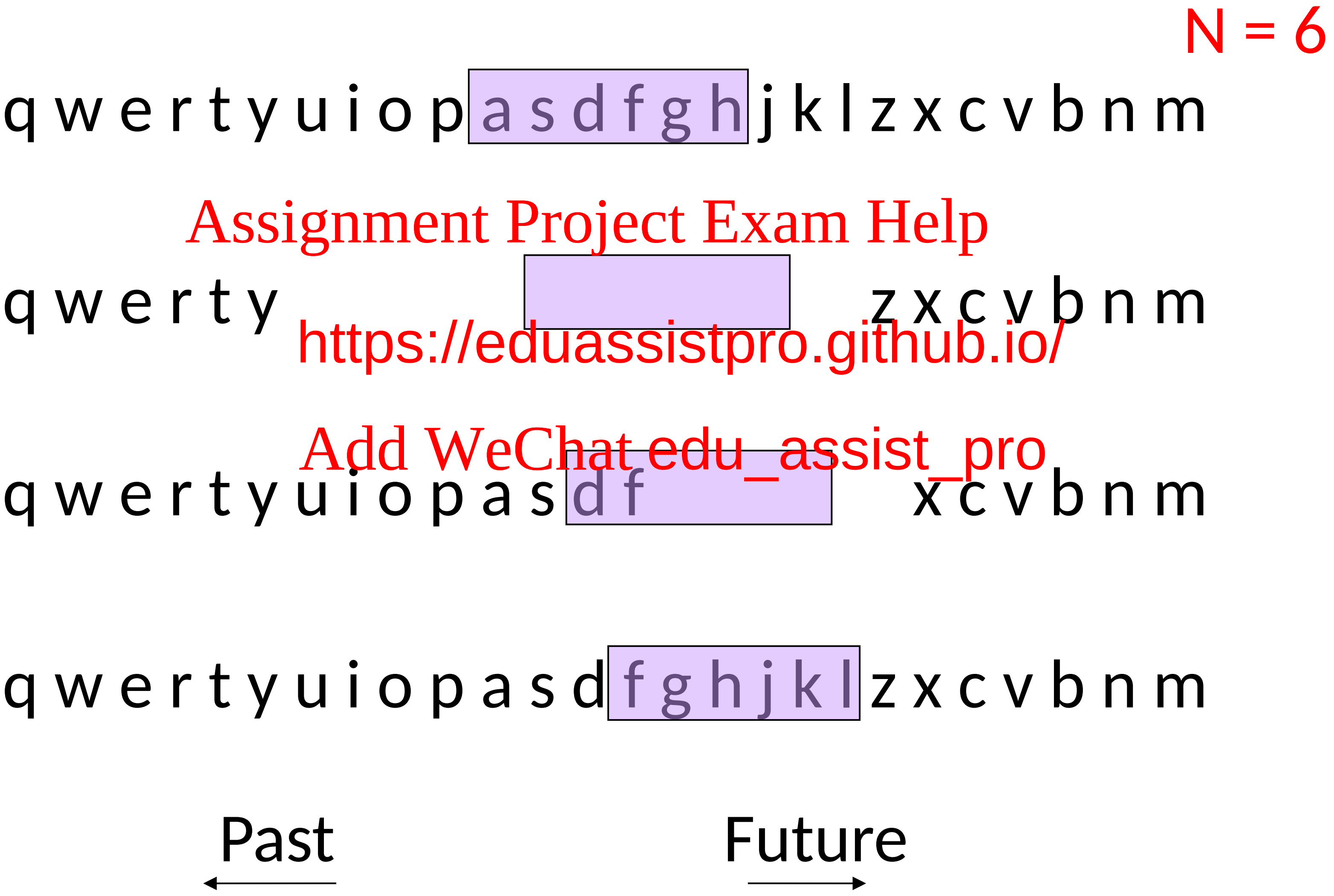
Outline

- Sampling from a Data Stream
- Queries over a (long) Sliding Windows
- Filtering Dat
- Counting Distinct Eleme
- Computing Moments
- Counting Itemsets

Sliding Windows

- A useful model of stream processing is that queries are about a *window* of length *N* – the *N* most rece ed
- Interesting c t cannot be stored in memory, or e sk
  - Or, there are so many streams that windows for all cannot be stored

A Sliding Window Example



N = 6

q w e r t y u i o p a s d f g h j k l z x c v b n m

q w e r t y z x c v b n m

q w e r t y u i o p a s d f x c v b n m

q w e r t y u i o p a s d f g h j k l z x c v b n m

Past

Future

Counting Bits (1)

- Problem:
  - Given a stream of 0s and 1s
  - Be prepare f the form How many 1’s in e $k \leq N$

- Obvious solution:
  - Store the most recent N bits
  - When a new bit comes in, discard the $N+1^{st}$ bit

Counting Bits (2)

- You cannot get an exact answer without storing the entire window

- Real Proble ot afford to store *N* bits?

  - E.g., we are processing 1 reams and *N* = 1 billion

- But we’re happy with an approximate answer

An Attempt: Simple Solution

- How many 1s are in the last *N* bits?
- Simple solution that does not really solve our problem: Uniformity ass
- Maintain 2 cou
  - *S*: number of 1s so far
  - *Z*: number of 0s so far
- How many 1s are in the last *N* bits? *N·S/(S+Z)*
- But, what if stream is non-uniform?
  - What if distribution changes over time?

[Datar, Gionis, Indyk, Motwani]

DGIM Method

- Store O(log^2N) bits per stream

- Gives approx ver off by more than 50%

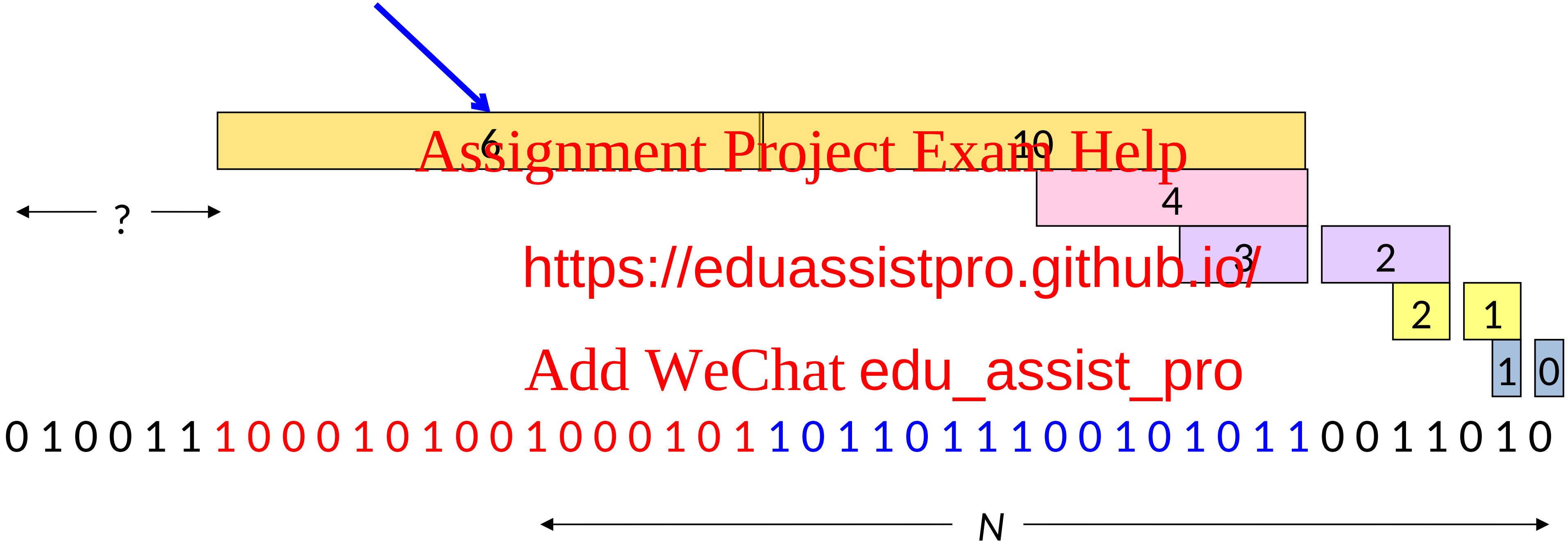
  - Error factor can be redu fraction > 0, with more complicated algorithm and proportionally more stored bits

Idea: Exponential Windows

- Solution that doesn’t (quite) work:
  - Summarize exponentially increasing regions of the stream, loo
  - Drop small r at the same point as a larger region

An Exponential Window Example

Window of width 16 has 6 1s



We can construct the count of the last *N* bits, except we're not sure how many of the last 6 are included.

What’s Good?

- Stores only O(log^2*N*) bits
  - O(log *N*) counts of log *N* bits each

- Easy update as more bit

- Error in count no greater than the number of 1s in the “unknown” area

What’s Not So Good?

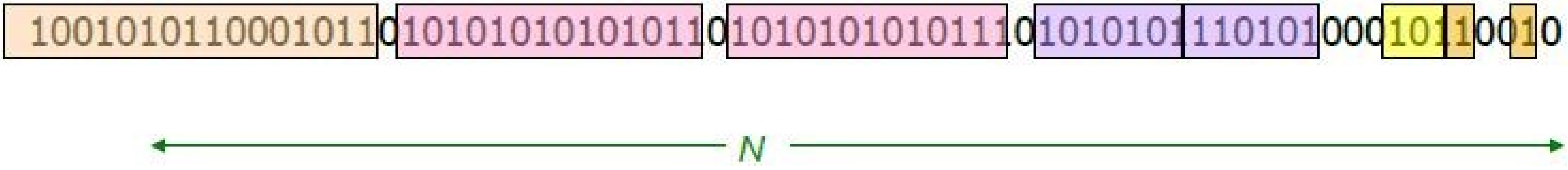
- As long as the 1s are fairly evenly distributed, the error ratio due to the unknown region is small – no m
- But it could re in the unknown area at the en
- In that case, the error is unbounded

[Datar, Gionis, Indyk, Motwani]

Fixup: DGIM Method

- Instead of summarizing fixed-length blocks, summarize blocks with specific numbers of 1s
  - Let the bloc) increase exponentiall

- When there are few 1s i dow, block sizes stay small, so errors are small



DGIM: Timestamps

- Each bit in the stream has a *timestamp*, starting 1, 2, ….
- Record times (the window size), so we c vant timestamp in O(log$_2$*N*)

DGIM: Buckets

- A *bucket* in the DGIM method is a record consisting of:

1. The times [illegible] g *N*) bits]

2. The numb [illegible] beginning and end: [O(log log *N*) bits

- Constraint on buckets: Number of 1s must be a power of 2
  - That explains the O(log log *N*) in (2)

Representing a Stream by Buckets

- Either one or two buckets with the same power-of-2 number of 1s

- Buckets do n stamps

- Buckets are sorted by si

  - Earlier buckets are not s n later buckets
- Buckets disappear when their end-time is > *N* time units in the past

Example: Bucketized Stream

Properties we maintain:

- Either one or two buckets with the same power-of-2 number of 1s
- Buckets do not overlap in timestamp
- Buckets are sorted by size

Updating Buckets – (1)

- When a new bit comes in, drop the last (oldest) bucket if its end-time is prior to *N* time units before the cu

- 2 cases: Current bit is 0 o

- If the current bit is 0, no other changes are needed

Updating Buckets – (2)

- If the current bit is 1:
  - Create a new bucket of size 1, for just this bit
    - End timest
  - If there are f size 1, combine the oldest two into a bu e 2
  - If there are now three buckets of size 2, combine the oldest two into a bucket of size 4
  - And so on …

Example

How to Query?

- To estimate the number of 1s in the most recent *N* bits:

  - Sum the siz the last

  - Add half the et

- Remember: we don't kn many 1s of the last bucket are still within the window

Example: Bucketized Stream

In-Class Practice 1

- Go to practice

Error Bound: Proof

- Suppose the last bucket has size 2^r
- Then by assuming 2^{r-1} of its 1s are still within the window, of at most 2^{r-1}
- Since there is at least o t of each of the sizes less than 2^r, th m is no less than $1 + 2 + 4 + \ldots + 2^{r-1} = 2^r - 1$
- Thus, error ratio is at most $2^{r-1} / (2^r - 1) \approx 50\%$

Extensions (For Thinking)

- Can we use the same trick to answer queries “How many 1s in the last *k*?” where *k* < *N*?

- Can we handle the case he stream is not bits, but integers, a nt the sum of the last *k*?

Reducing the Error

- Instead of maintaining 1 or 2 of each size bucket, we allow either *r*-1 or *r* for *r* > 2
  - Except for t [illegible] ts; we can have any number [illegible] those
- Error is at most 1/*r*
- By picking r appropriately, we can tradeoff between number of bits and the error

Outline

- Sampling from a Data Stream
- Queries over a (long) Sliding Windows
- Filtering Dat
- Counting Distinct Eleme
- Computing Moments
- Counting Itemsets

Filtering Data Streams

- Each element of data stream is a tuple (a finite list of elements)
- Given a list of keys *S*

- How to deter ts of stream have keys in *S*?

- Obvious solution: Hash table
  - But suppose we do not have enough memory to store all of *S* in a hash table
    - E.g., we might be processing millions of filters on the same stream

Applications

- Example: Email spam filtering
  - We know 1 billion “good” email addresses
  - If an email c ese, it is NOT spam
- Publish-subscribe syste
  - People express interest in certain sets of keywords
  - Determine whether each message matches user’s interest

First Cut Solution – (1)

- Given a set of keys *S* that we want filter
- Create a bit array *B* of *n* bits, initially all 0s

- Choose a has range *[0,n)*

- Hash each m e of ts, and set that bit to 1, i.e. *1*

- Hash each element *a* of the stream and output only those that hash to bit that was set to 1
  - Output a if *B[h(a)] == 1*

First Cut Solution – (2)

- Creates false positives but no false negatives
  - If the item is in *S* we surely output it, if not we may still output it

First Cut Solution – (3)

- $|S|$ = 1 billion email addresses

  $|B|$ = 1GB = 8 billion bits
- If the email a it surely hashes to a bucked t o 1, so it always gets through (no false

- Approximately 1/8 of the bits are set to 1, so about 1/8th of the addresses not in S get through to the output (false positives)
  - Actually, less than 1/8th, because more than one address might hash to the same bit

Analysis: Throwing Darts

- More accurate analysis for the number of false positives

- Consider: If nto *n* equally likely targets bility that a target gets at least one

- In our case:
  - Targets = bits/buckets
  - Darts = hash values of items

Analysis: Throwing Darts – (2)

- We have *m* darts, *n* targets
- What is the probability that a target gets at least one dar

Analysis: Throwing Darts – (3)

- Fraction of 1s in the array *B* == probability of false positive ==

- Example: dar
  - Fraction of 1s in B = = 0.1
  - Compare with our earlier estimate: 1/8 = 0.125

- Can we improve this error?

Bloom Filter

- Consider: $|S| = m$, $|B| = n$
- Use k independent hash functions

- Initialization:
  - Set B to all 0

  - Hash each element using h function , set (for each)
- Run-time:
  - When a stream element with key *x* arrives
    - If for all , then declare that *x* is in *S*
    - Otherwise discard the element *x*

Bloom Filter – Analysis

- What fraction of the bit vector *B* are 1s?
  - Throwing *k*·*m* darts at *n* targets
  - So fraction
- But we have sh functions
- So, false positive proba

Bloom Filter – Analysis (2)

- *m* = 1 billion, *n* = 8 billion
  - k = 1: = 0.1175
  - k = 2: = 0.0

- What happens as we ke increasing *k*?
- “Optimal” value of k:
  - E.g.:

Bloom Filter: Wrap-up

- Bloom filters guarantee no false negatives, and use limited memory
  - Great for pr more expensive checks
  - E.g., Google's BigTable, proxy
- Suitable for hardware implementation
  - Hash function computations can be parallelized

Outline

- Sampling from a Data Stream
- Queries over a (long) Sliding Windows
- Filtering Dat
- Counting Distinct Eleme
- Computing Moments
- Counting Itemsets

Counting Distinct Elements

- Problem:
  - Data stream consists of a universe of elements chosen fro
  - Maintain a of distinct elements seen so far
- Obvious approach:
  - Maintain the set of elements seen so far

Applications

- How many different words are found among the Web pages being crawled at a site?

  - Unusually lo ould indicate artificial pag

- How many different Web pages does each customer request in a week?

Using Small Storage

- Real Problem: What if we do not have space to store the complete set?

- Estimate the count in a d way

- Accept that the count may be in error, but limit the probability that the error is large

Flajolet-Martin Approach

- Pick a hash function *h* that maps each of the *n* elements to at least $\log_2 N$ bits

- For each stre *a*) be the

number of tra

  - *r*(*a*) = position of first 1 co m the right

- Record *R* = the maximum *r*(*a*) seen
  - $R = \max_a r(a)$, over all the items a seen so far

- Estimated number of distinct elements = 2^R

Based on a variant due to AMS (Alon, Matias, and Szegedy)

Why It Works

- The probability that a given $h(a)$ ends in at least r 0s is 2^{-r}
  - – $h(a)$ hashe y at random
  - – Probabilit ber ends in at least r 0s is 2^{-r}
- If there are m different elements, the probability that $R \geq r$ is $1 - \boxed{\boxed{(1 - 2^{-r})}^m}$

Prob. all $h(a)$'s end in fewer than r 0s.

Prob. a given $h(a)$ ends in fewer than r 0s.

Why It Works – (2)

- Note:
- Prob. of NOT finding a tail of length *r* is:

  - If , then prob.
    - as

    - So, the probability of find length *r* tends to 0

  - If , then prob. tends to 0
    - as
    - So, the probability of finding a tail of length *r* tends to 1
- Thus, will almost always be around *m*.

Why It Doesn’t Work

- E[2^R] is actually infinite
  - Probability halves when R R +1, but value doubles
- Workaround involves using many hash functions and getting m

- How are samples combine

  - Average? What if one very large value?
  - Median? All values are a power of 2
  - Solution:
    - Partition your samples into small groups
    - Take the average of groups
    - Then take the median of the averages

In-Class Practice 2

- Go to practice

One-Slide Takeaway

- Sampling from a streaming data
  - How to get a fixed proportion or a fixed-size Sample

- Queries over a l
  - understand DG

- Filtering Data Streams

  - understand first cut solution and Bloom Filter
- Counting distinct elements
  - Understand Flajolet-Martin Approach
- Appendix: computing moments and counting item sets

References

- Book:
  - Mining of Massive Datasets

- Massive Online Analysis oftware:
  - http://moa.cms.waikato

Appendix

- Sampling from a Data Stream
- Queries over a (long) Sliding Windows
- Filtering Dat
- Counting Distinct Eleme
- **Computing Moments**
- Counting Itemsets

Generalization: Moments

- Suppose a stream has elements chosen from a set of *N* values

- Let m_a be the number o alue *a* occurs

- The k^{th} *moment* is

Special Cases

- 0th moment = number of different elements
  - The problem just considered
- 1st moment = bers of elements = l
  - Easy to compute

- 2nd moment = *surprise number* = a measure of how uneven the distribution is

Example: Surprise Number

- Stream of length 100; 11 values appear
- Item counts: 10, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9

  Surprise # =
- Item counts: 90, 1, 1, 1, ,1, 1, 1

  Surprise # = 8,110

Due to AMS (Alon, Matias, and Szegedy)

AMS Method

- Works for all moments
- Gives an unbiased estimate

- We’ll just con ment

- Based on calculation of many random variables *X*:
  - For each random variable *X*, we store *X.el* and *X.val*
  - Each random variable represents one separate item
  - Note this requires a count in main memory, so number of *X*s is limited

One Random Variable

- Assume stream has length *n*
- Pick a random time to start, so that any time is equally likely

- Let the chose nt *a* in the stream

- *X* = *n* * ((twice the number of *a*s in the stream starting at the chosen time) – 1)
  - Note: store *n* once, count of *a*s for each *X*

Expected Value of X

- 2nd moment is
- E(X) = $_{\text{all times } t}$ n * ([twice the number of times the stream e ppears from that time on] – 1)

=

=

Group times by the value seen

Time when the last *a* is seen

Time when the penultimate *a* is seen

Time when the first *a* is seen

Combining Samples

- One random variable only represent one sampled item; we should do many concurrent samples
- Compute as many variables X as can fit in available memory
- Average them in groups
- Take median of averages
- Proper balance of group sizes and number of groups assures not only correct expected value, but expected error goes to 0 as number of samples gets large

Problem: Streams Never End

- We assumed there was a number *n*, the number of positions in the stream
- But real stre , so *n* is a variable – th s seen so far

Stream Never End: Fixups

- The variables *X* have *n* as a factor – keep *n* separately; just hold the count in *X*
- Suppose we can only store *k* counts. We must throw some *X* 's out
  - Objective: each starting tim ed with probability *k*/*n*
  - Solution: (fix-size sampling!)
    - Choose the first *k* times for *k* variables
    - When the n^{th} element arrives ($n > k$), choose it with probability *k*/*n*
    - If you choose it, throw one of the previously stored variables out, with equal probability

Appendix

- Sampling from a Data Stream
- Queries over a (long) Sliding Windows
- Filtering Dat
- Counting Distinct Eleme
- Computing Moments
- Counting Itemsets

Counting Itemsets

- New Problem: Given a stream, which items appear more than *s* times in the window?

- Possible solu stream of baskets as o er item

  - 1 = item present; 0 = not
  - Use DGIM to estimate counts of 1s for all items

Extensions

- In principle, you could count frequent pairs or even larger sets the same way
  - One stream

- Drawbacks:
  - Only approximate
  - Number of itemsets is way too big

Exponentially Decaying Windows

- Exponentially decaying windows: A heuristic for selecting likely frequent itemsets

  - What are “c lar movies?
    - Instead of t in last *N* elements

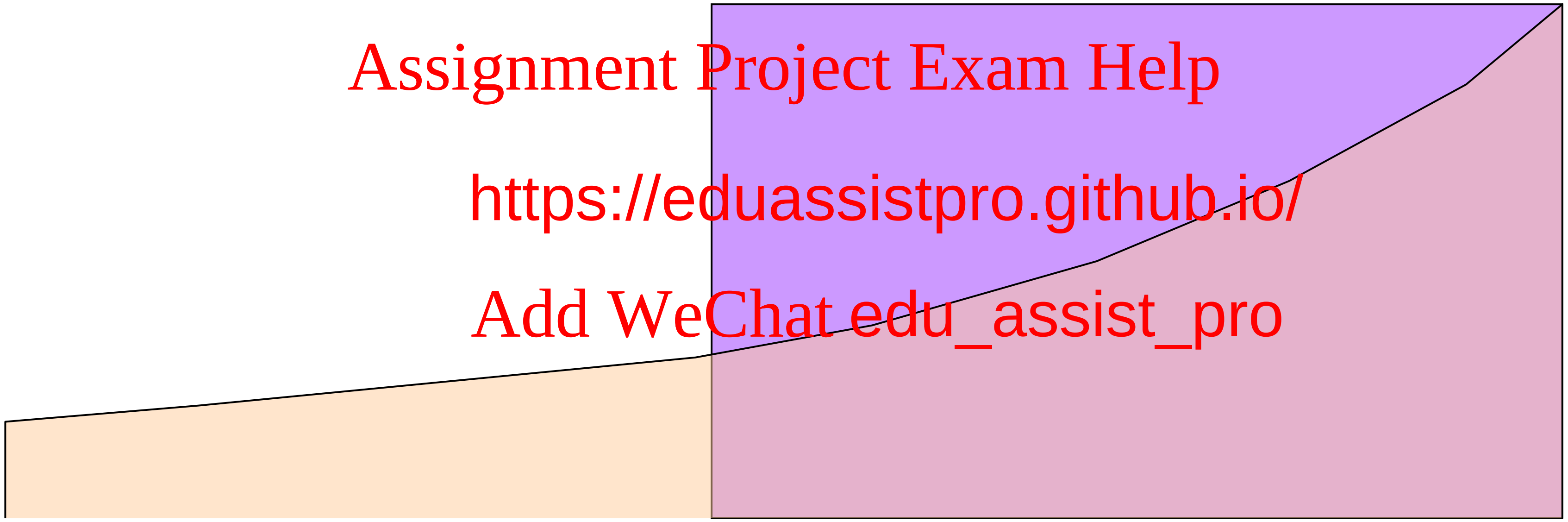
    - Compute a smooth aggre the whole stream

- If stream is a_1, a_2,… and we are taking the sum of the stream, take the answer at time *t* to be:
  - *c* is a constant, presumably tiny, like or
  - When new arrives: Multiply current sum by *(1-c)* and add

Example: Counting Items

- If each is an "item" we can compute the characteristic function of each possible item *x* as an exponentially decaying window (E.D.W.).
  - That is:
    - where if , and 0 otherwise
  - Imagine that for each item a binary stream (1 … *x* appears, 0 … *x* does not appear)
  - New item *x* arrives:
    - Multiply all counts by *(1-c)*
    - Add +1 to count for *x*
- Call this sum the "weight" of item *x*

Sliding Versus Decaying Windows



Important property: Sum over all weights is

=

Counting Items

- Suppose we want to find those items of weight > ½

  - Important p l weights is =

- Thus:

  - There cannot be more t with weight of ½ or more
- So, is a limit on the number of movies being counted at any time

Extension to Larger Itemsets

- Count (some) itemsets in an E.D.W.
  - Problem: Too many itemsets to keep counts of all of them in memory

- When a bask

  - Multiply all counts by (1-

  - For uncounted items in *B*, create new count
  - Add 1 to count of any item in *B* and to any itemset contained in *B* that is already being counted
  - Drop counts < ½
  - Initiate new counts (next slide)

Initiation of New Counts

- Start a count for an itemset if every proper subset of *S* had a count prior to arrival of basket *B*
  - Intuitively: If all subsets of S are being counted this means they a thus S has a potential to be “hot”

- Example
  - Start counting {*i*, *j*} iff both *i* and *j* were counted prior to seeing *B*
  - Start counting {*i*, *j*, *k*} iff {*i*, *j*}, {*i*, *k*}, and {*j*, *k*} were all counted prior to seeing *B*

How Many Counts?

- Counts for single items < (2/*c*) * (average number of items in a basket)
- Counts for la
- But we are conservative tarting counts of large sets
  - If we counted every set we saw, one basket of 20 items would initiate 1M counts

In-Class Practice 1

- There are several ways that the bit-stream 1001011011101 could be partitioned into buckets. Fin

In-Class Practice 2

- Suppose our stream consists of the integers 3, 1, 4, 1, 5, 9, 2, 6, 5. Our hash functions will all be of the form for some *a* and *b*. You should treat the result as a 5-bit binary inte ail length for each stream element and the re timate of the number of distinct element sh function is: