Processing Data Streams

Andrew McGregor
University of Pennsylvania

Data Stream Model

[Morris '78] [Munro, Paterson '78] [Flajolet, Martin '85] [Alon, Matias, Szegedy '96]
[Henzinger, Raghavan, Rajagopalan '98] [Feigenbaum, Kannan, Strauss, Viswanathan '99]

Data Stream Model

[Morris '78] [Munro, Paterson '78] [Flajolet, Martin '85] [Alon, Matias, Szegedy '96]
[Henzinger, Raghavan, Rajagopalan '98] [Feigenbaum, Kannan, Strauss, Viswanathan '99]

Data Stream Model

[Morris '78] [Munro, Paterson '78] [Flajolet, Martin '85] [Alon, Matias, Szegedy '96]
[Henzinger, Raghavan, Rajagopalan '98] [Feigenbaum, Kannan, Strauss, Viswanathan '99]

- Constraints: Sequential access to data
  Small working memory
  Fast processing of elements

Data Stream Model

[Morris '78] [Munro, Paterson '78] [Flajolet, Martin '85] [Alon, Matias, Szegedy '96]
[Henzinger, Raghavan, Rajagopalan '98] [Feigenbaum, Kannan, Strauss, Viswanathan '99]



- Constraints: Sequential access to data
  Small working memory
  Fast processing of elements


Streams

Single Pass

e.g. network monitoring

Multi-Pass

e.g. huge log files

PASSES

Poly-log
e.g. statistics

Sub-linear
e.g. functions

SPACE

Single Pass
e.g. network monitoring

Multi-Pass
e.g. huge log files

PASSES

Time Series
e.g. stock prices

Adversarial
e.g. safety-critical

ORDER

Poly-log
e.g. statistics

Sub-linear
e.g. functions

SPACE

Single Pass
e.g. network monitoring

Multi-Pass
e.g. huge log files

PASSES

Time Series
e.g. stock prices

Adversarial
e.g. safety-critical

ORDER

Poly-log
e.g. statistics

Sub-linear
e.g. functions

SPACE

Single Pass
e.g. network monitoring

Multi-Pass
e.g. huge log files

PASSES

Numerical Data
e.g. quantiles, distances

Geometric Data
e.g. clusters, convex hulls

DATA

Does it matter how streams are ordered?

ORDER

Poly-log
e.g. statistics

Sub-linear
e.g. functions

SPACE

Single Pass
e.g. network monitoring

Multi-Pass
e.g. huge log files

PASSES

Numerical Data
e.g. quantiles, distances

Geometric Data
e.g. clusters, convex hulls

DATA

Does it matter how streams are ordered?

ORDER

Poly-log
e.g. statistics

Sub-linear
e.g. functions

SPACE

How valuable is an extra pass over the data?

PASSES

Numerical Data
e.g. quantiles, distances

Geometric Data
e.g. clusters, convex hulls

DATA

Does it matter how streams are ordered?

ORDER

Can we create small-space sketches?

SPACE

How valuable is an extra pass over the data?

PASSES

Numerical Data

e.g. quantiles, distances

Geometric Data

e.g. clusters, convex hulls

DATA

Does it matter how streams are ordered?

ORDER

Can we create small-space sketches?

SPACE

How valuable is an extra pass over the data?

PASSES

What about a stochastic-stream or graph-stream?

DATA

1. Thesis Overview
2. Some Algorithms
3. Some Lower-Bounds

I. Thesis Overview

Questions

Stochastic Streams

Graph Streams

Overview of Chapters

Questions

Questions

- Does it matter how streams are ordered?

Questions

- Does it matter how streams are ordered?

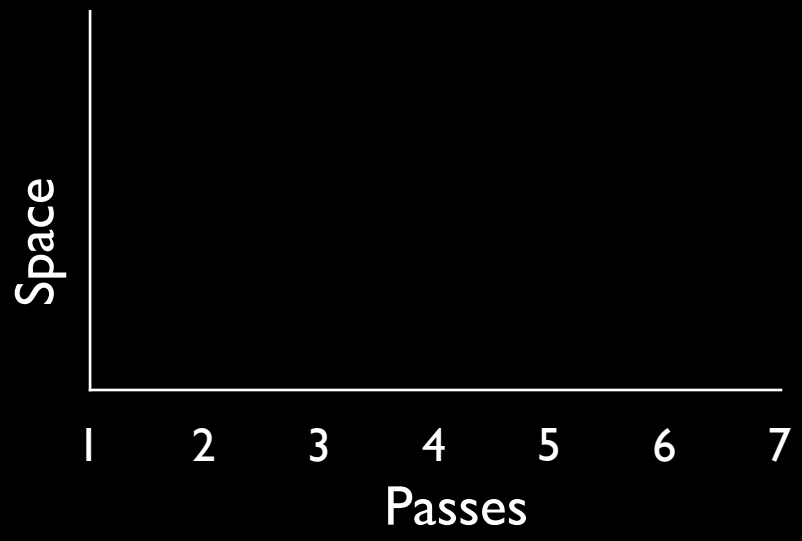
Yes!

Questions

- Does it matter how streams are ordered?
- How valuable is are extra pass over the data?

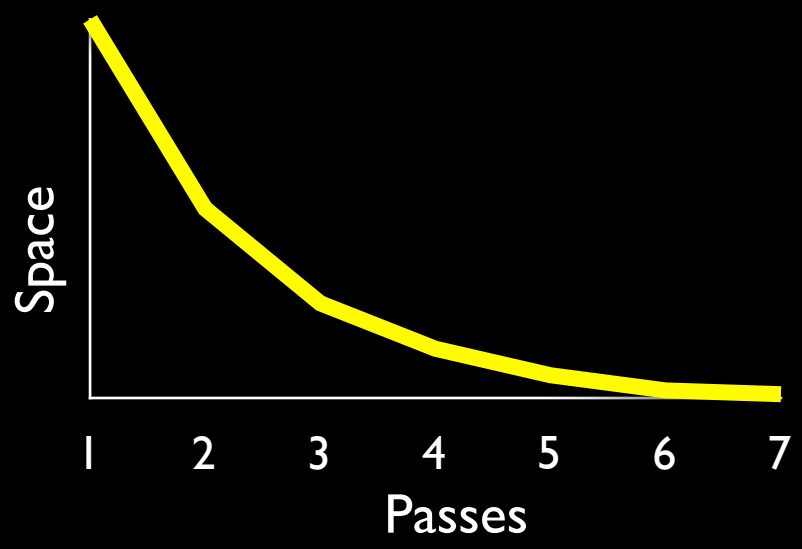
Questions

- Does it matter how streams are ordered?
- How valuable is are extra pass over the data?



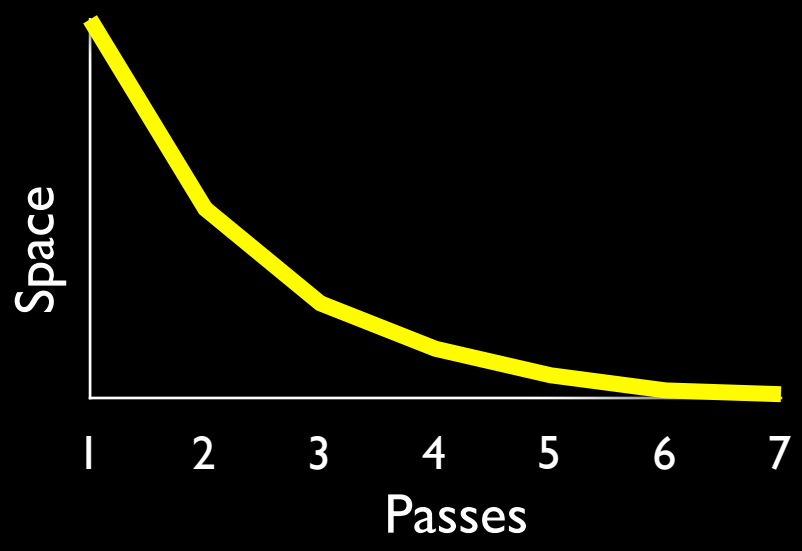
Questions

- Does it matter how streams are ordered?
- How valuable is are extra pass over the data?

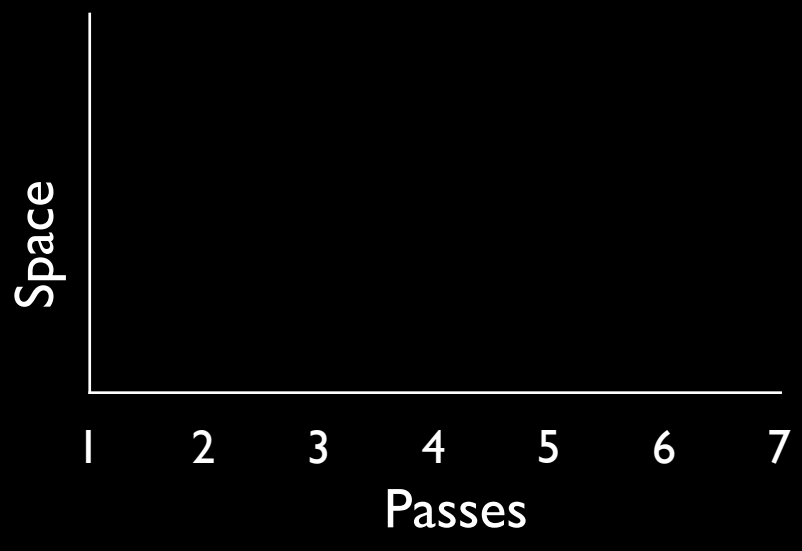


Questions

- Does it matter how streams are ordered?
- How valuable is are extra pass over the data?

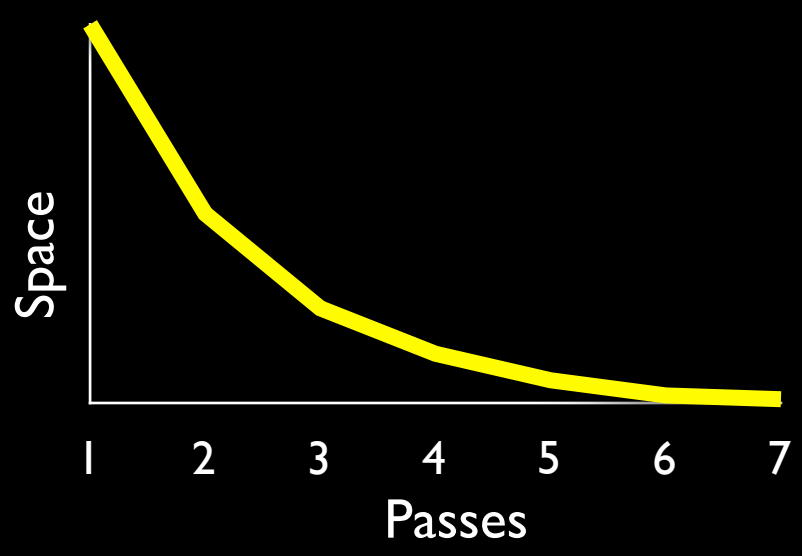


?

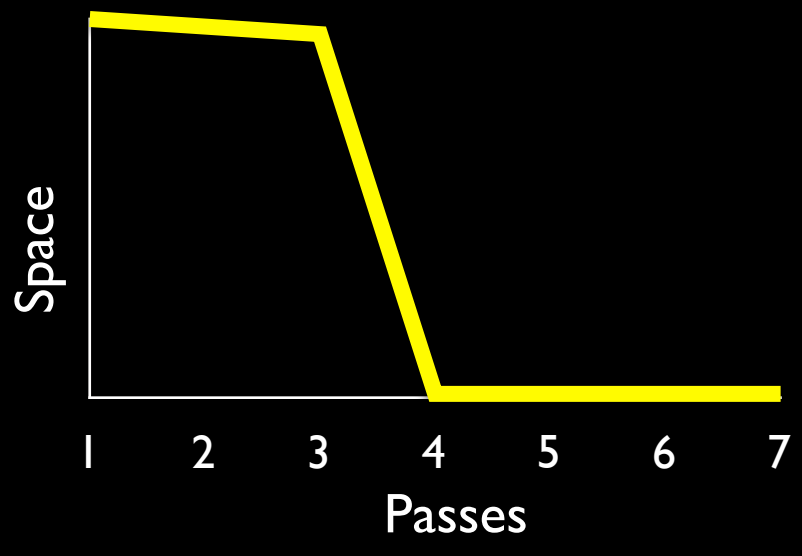


Questions

- Does it matter how streams are ordered?
- How valuable is are extra pass over the data?



?



Questions

- Does it matter how streams are ordered?
- How valuable is are extra pass over the data?
- Can we create small-space sketches?

Questions

- Does it matter how streams are ordered?
- How valuable is are extra pass over the data?
- Can we create small-space sketches?

New Stream

Questions

- Does it matter how streams are ordered?
- How valuable is are extra pass over the data?
- Can we create small-space sketches?

Sketch of Old Stream

Can estimate L_1 distance.
[Indyk '00]

New Stream

Questions

- Does it matter how streams are ordered?
- How valuable is are extra pass over the data?
- Can we create small-space sketches?

Sketch of Old Stream

Can estimate L_1 distance.
[Indyk '00]

What about the relative distance or Hellinger distance?

New Stream

Questions

- Does it matter how streams are ordered?
- How valuable is are extra pass over the data?
- Can we create small-space sketches?

Sketch of Old Stream

Can estimate L_1 distance.
[Indyk '00]

What about the relative distance or Hellinger distance?

No!

New Stream

Stochastic Streams

Stochastic Streams

- Stream: $a_1, a_2, \ldots, a_m$ where $a_j \in [n]$

  Frequency: m_i = number of occurrences of i

  Empirical Probability: $p_i = m_i/m$

Stochastic Streams

- Stream: $a_1, a_2, \ldots, a_m$ where $a_j \in [n]$

  Frequency: m_i = number of occurrences of i

  Empirical Probability: $p_i = m_i/m$

- Compute some function of $(p_1, p_2, \ldots, p_n)$

Stochastic Streams

- Stream: $a_1, a_2, \ldots, a_m$ where $a_j \in [n]$

  Frequency: m_i = number of occurrences of i

  Empirical Probability: $p_i = m_i/m$

- Compute some function of $(p_1, p_2, \ldots, p_n)$

e.g. $\sum p_i^k$ [Alon, Matias, Szegedy '96]

$\sum |p_i - q_i|$ [Indyk '00]

$\sum_{i \le j} p_i \approx 1/2$ [Greenwald, Khanna '01]

Stochastic Streams

- Stream: $a_1, a_2, \ldots, a_m$ where $a_j \in [n]$

  Frequency: m_i = number of occurrences of i

  Empirical Probability: $p_i = m_i/m$

- Compute some function of $(p_1, p_2, \ldots, p_n)$

e.g. $\sum -p_i \log p_i$ Entropy?

$\sum p_i \log(p_i/q_i)$ Kullback-Leibler?

$\sum (\sqrt{p_i} - \sqrt{q_i})^2$ Hellinger?

REAL Stochastic Streams

- Who cares about the empirical distribution?!

REAL Stochastic Streams

- Who cares about the empirical distribution?!
- What can we infer about the stream's source?

REAL Stochastic Streams

- Who cares about the empirical distribution?!
- What can we infer about the stream's source?
- Assume the stream consists of independent samples or an evolving Markov chain.

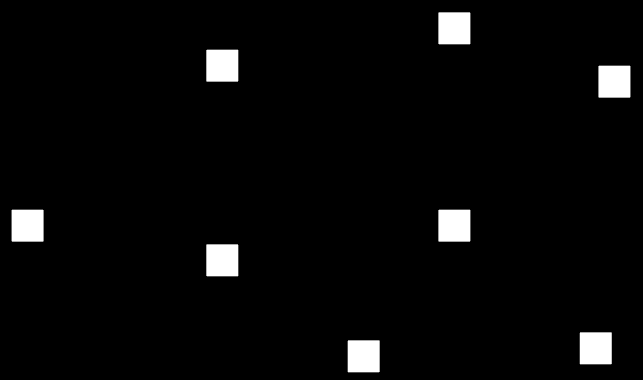
REAL Stochastic Streams

- Who cares about the empirical distribution?!
- What can we infer about the stream's source?
- Assume the stream consists of independent samples or an evolving Markov chain.
- Can we learn in small-space?

Graph Streams

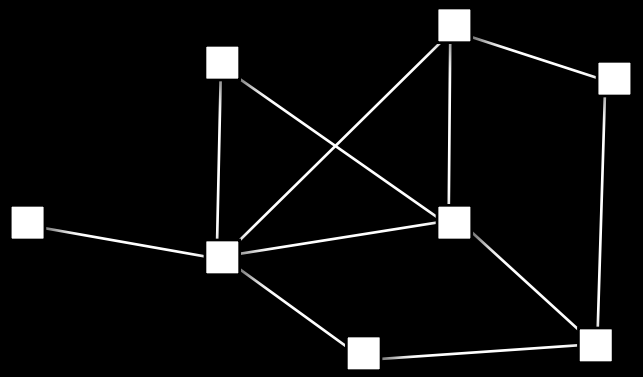
Graph Streams

- Stream: $a_1, a_2, \ldots, a_m$ where $a_j \in [n]^2$



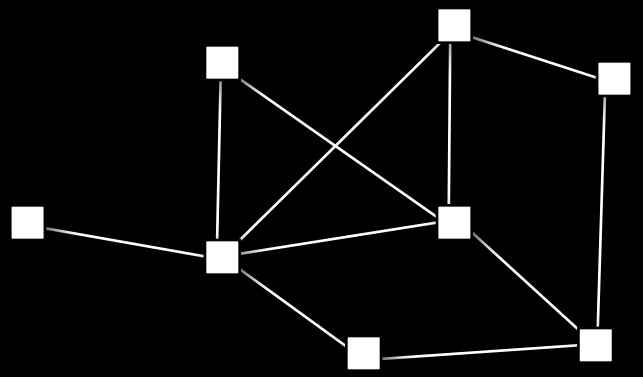
Graph Streams

- Stream: $a_1, a_2, \ldots, a_m$ where $a_j \in [n]^2$



Graph Streams

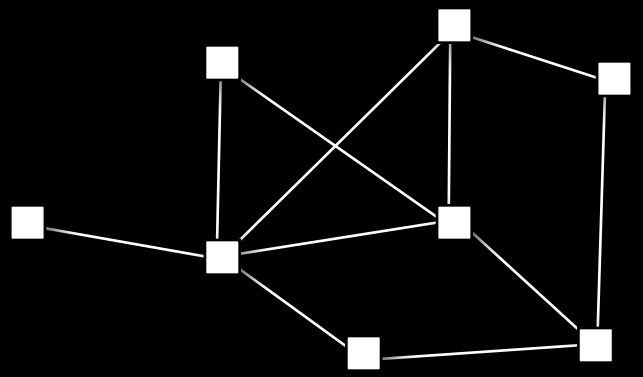
- Stream: $a_1, a_2, ..., a_m$ where $a_j \in [n]^2$



- E.g. Web-graph or Call-graph

Graph Streams

- Stream: $a_1, a_2, ..., a_m$ where $a_j \in [n]^2$

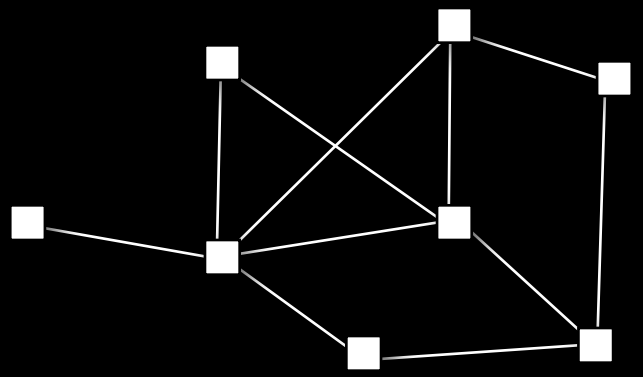


- E.g. Web-graph or Call-graph
- Previously: Triangles and Common Neighborhoods

  [Bar-Yossef, Kumar, Sivakumar '02] [Buchsbaum, Giancarlo, Westbrook '03]

Graph Streams

- Stream: $a_1, a_2, \ldots, a_m$ where $a_j \in [n]^2$



- E.g. Web-graph or Call-graph
- Previously: Triangles and Common Neighborhoods
  [Bar-Yossef, Kumar, Sivakumar '02] [Buchsbaum, Giancarlo, Westbrook '03]
- Connectivity, diameter, girth, matchings?

I. Oracles & Ordering

Does it matter how streams are ordered?

Quantiles
Histograms
Frequency Moments

Guha, M. (06)
Guha, M. (PODS 06)
Guha, M., Venkatasubramanian (SODA 06)

I. Oracles & Ordering

Does it matter how streams are ordered?

Quantiles
Histograms
Frequency Moments

Guha, M. (06)
Guha, M. (PODS 06)
Guha, M., Venkatasubramanian (SODA 06)

II. Entropy & Distances

What distances are "sketchable"?

Entropy
f-Divergences
Bregman-Divergences

Guha, M., Indyk (07)
Chakrabarti, Cormode, M. (SODA 07)

I. Oracles & Ordering

Does it matter how streams are ordered?

Quantiles
Histograms
Frequency Moments

Guha, M. (06)
Guha, M. (PODS 06)
Guha, M., Venkatasubramanian (SODA 06)

II. Entropy & Distances

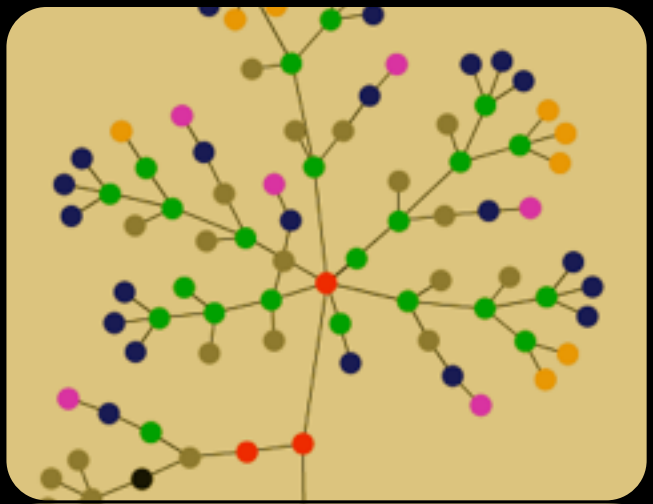
What distances are "sketchable"?

Entropy
f-Divergences
Bregman-Divergences

Guha, M., Indyk (07)
Chakrabarti, Cormode, M. (SODA 07)

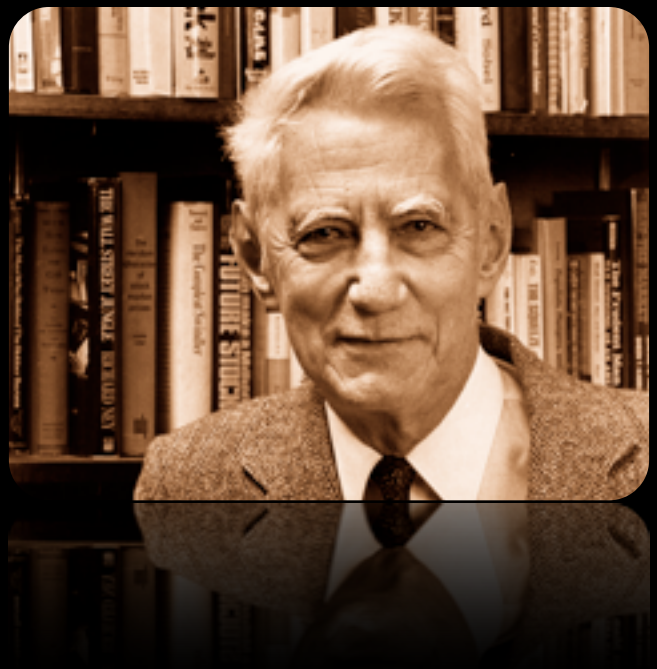
III. Graph Streams

How valuable is an extra pass?

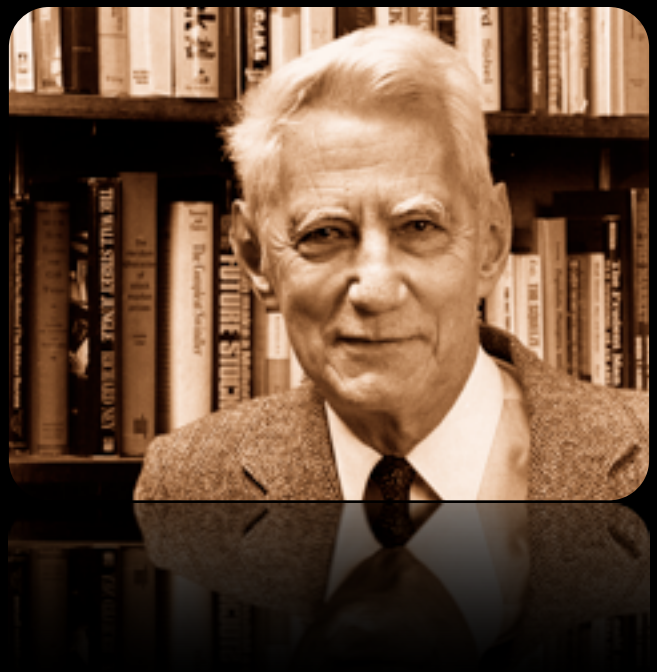


Girth
Diameter
Large Matchings

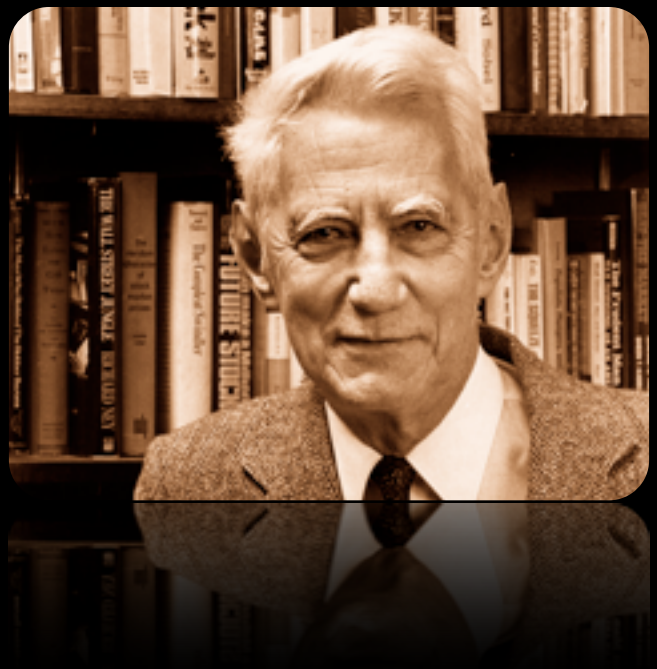
M. (APPROX 05)
Feigenbaum et al. (SODA 05)
Feigenbaum et al. (ICALP 04)


THE MULTISCIENCE LAW
FUTURE STUDIES

II. Algorithms for Estimating Entropy



II. Algorithms for Estimating Entropy



III. Use Entropy for Lower Bounds

1. Thesis Overview
2. Some Algorithms
3. Some-Lower Bounds

1. Thesis Overview

2. Some Algorithms

Estimating the entropy of a data stream

Higher-order entropy and randon-walks

Entropy

- Entropy: $H(p) = \sum -p_i \lg p_i$

  Upper-bound: $O(\varepsilon^{-2} \log m \log \delta^{-1})$

  Lower-bound: $\Omega(\varepsilon^{-2} / \log^2 \varepsilon^{-1})$

  [Guha et al. '06], [Chakrabarti et al. '06],
  [Lall et al. 06], [Bhuvanagiri, Ganguly '06].

- Generalizations:

  Higher-order entropy

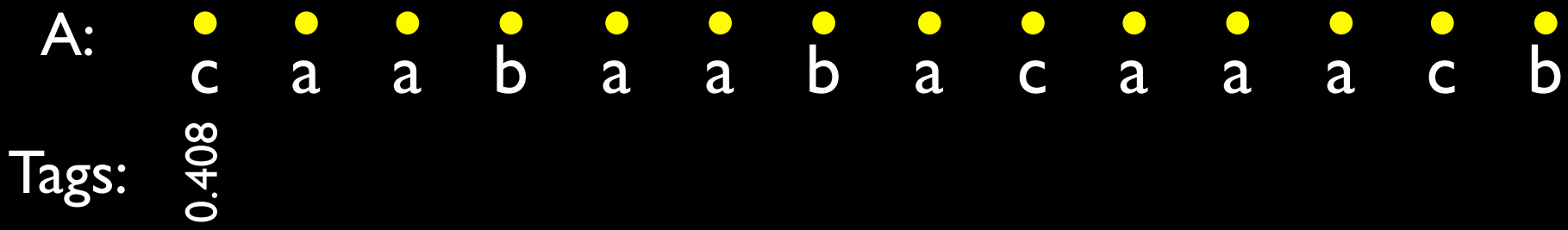
  Random-walk entropy


A:
c a a b a a b a c a a a c b

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$

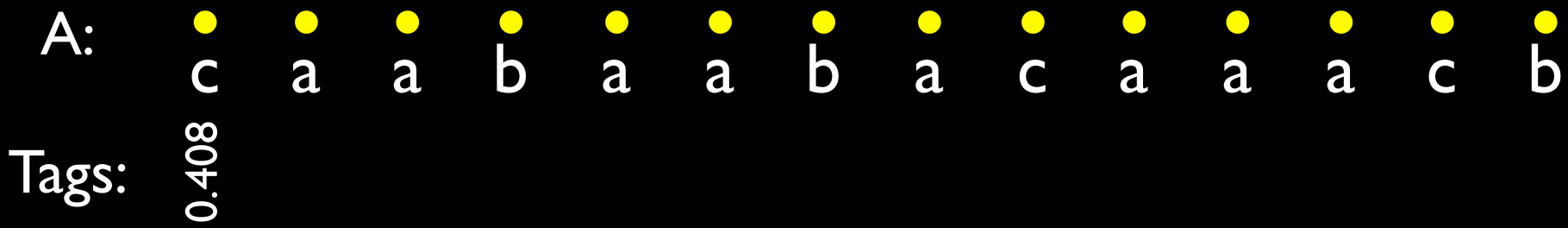
- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

$$i \in_R A \text{ and } j \in_R [m_i]$$



- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

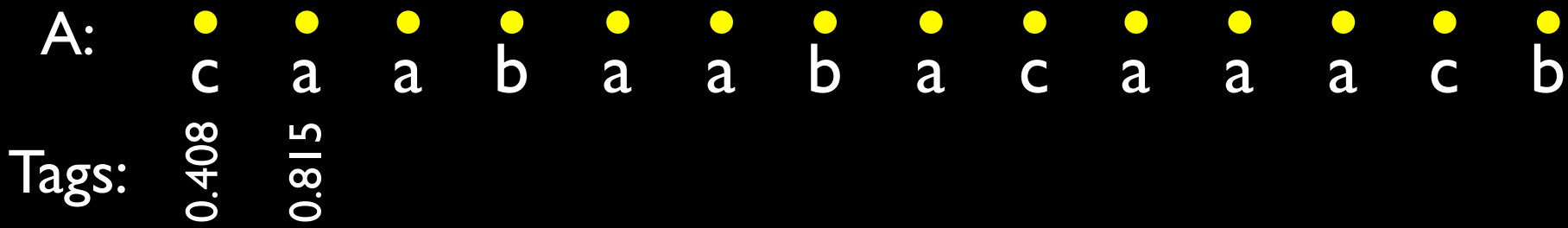
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.408, c, 1)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

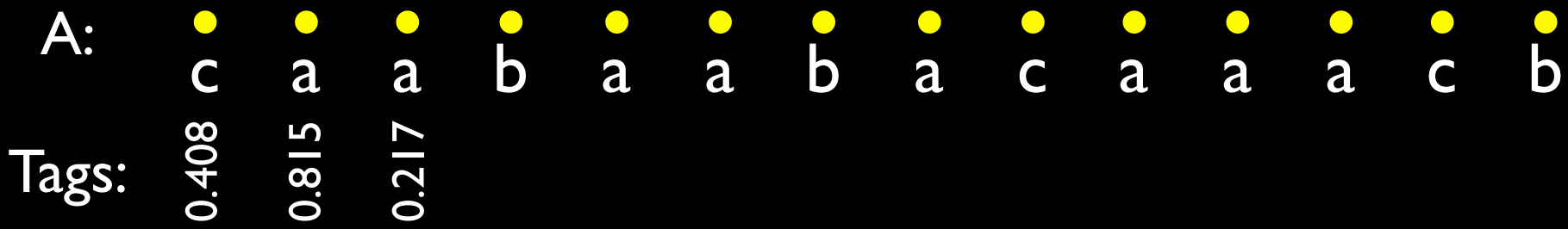
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.408, c, 1)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

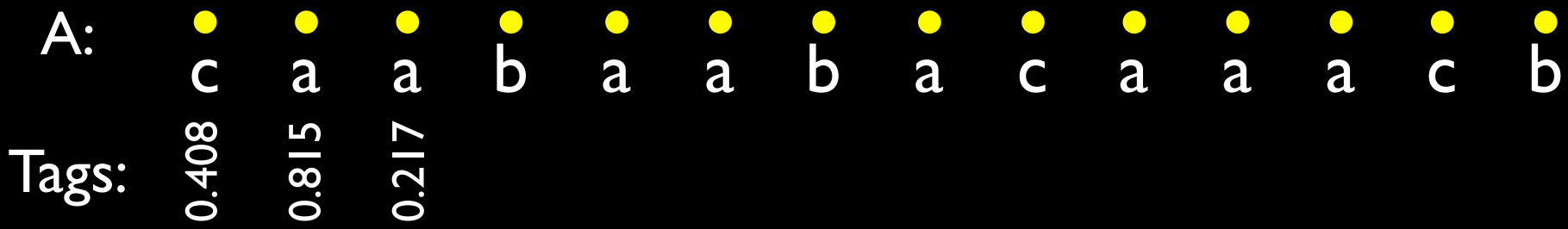
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.408, c, 1)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

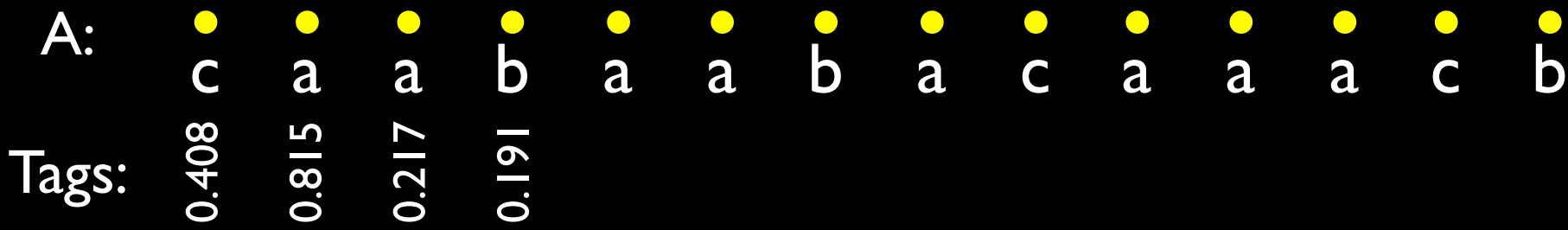
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.217, a, 1)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

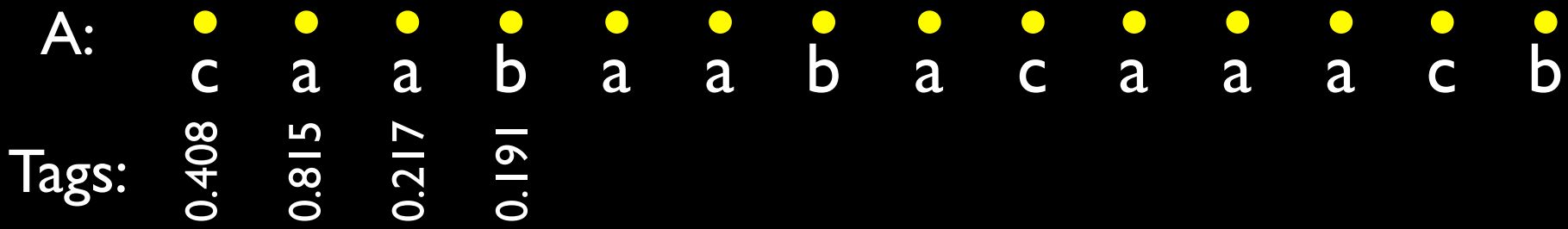
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.217, a, 1)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

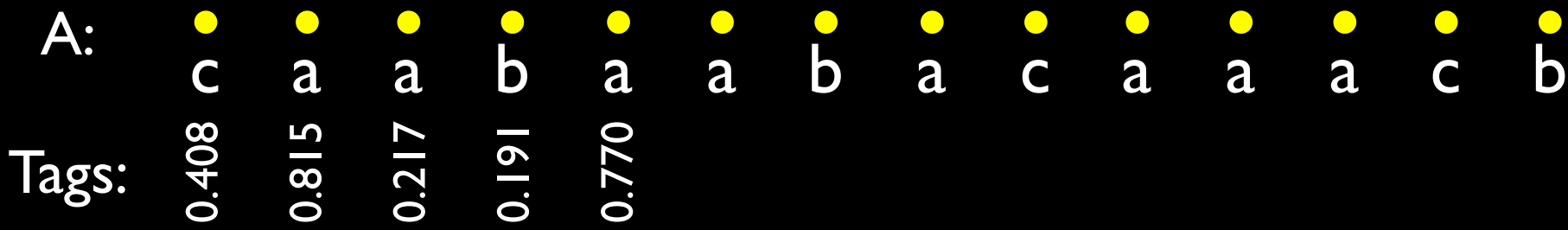
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.191, b, 1)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

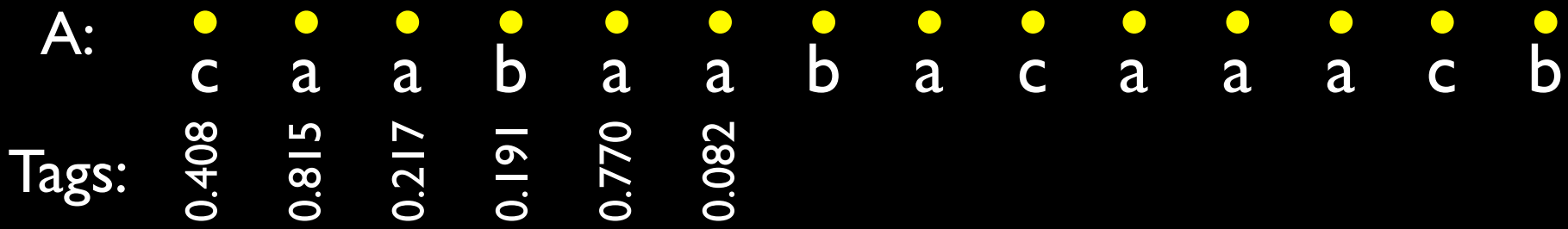
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.191, b, 1)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

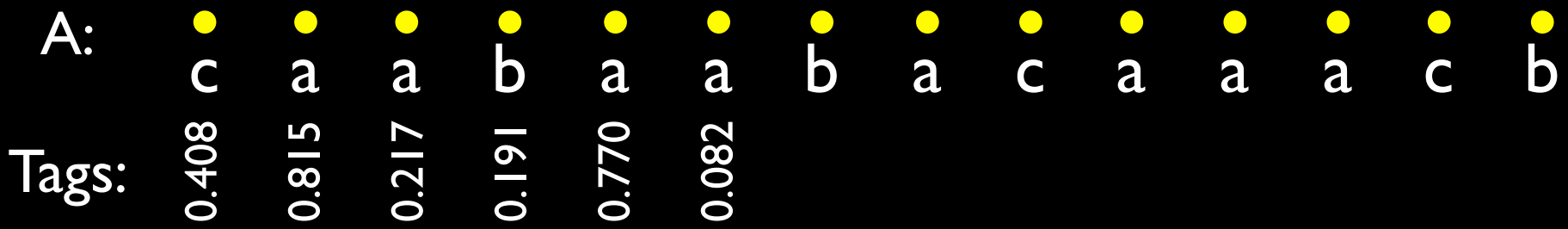
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.191, b, 1)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.082, a, 1)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

$$i \in_R A \text{ and } j \in_R [m_i]$$

A:	c	a	a	b	a	a	b	a	c	a	a	a	c	b
Tags:	0.408	0.815	0.217	0.191	0.770	0.082	0.366							

(min-tag, item, repeats)= (0.082, a, 1)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

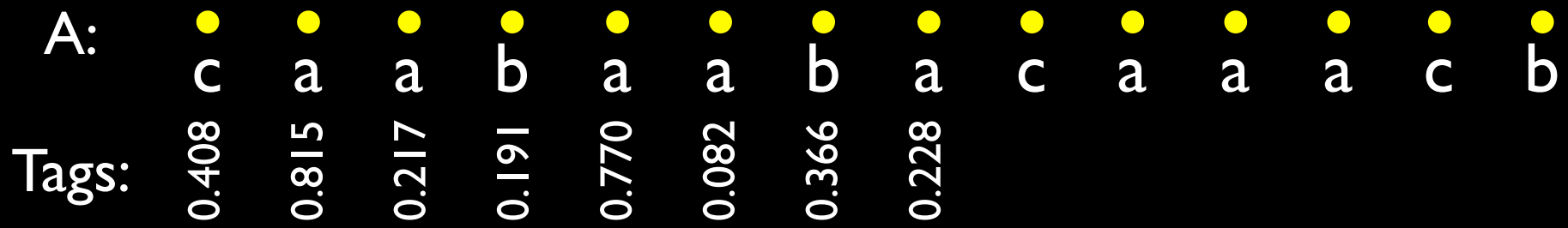
$$i \in_R A \text{ and } j \in_R [m_i]$$

A:	c	a	a	b	a	a	b	a	c	a	a	a	c	b
Tags:	0.408	0.815	0.217	0.191	0.770	0.082	0.366	0.228						

(min-tag, item, repeats)= (0.082, a, 1)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

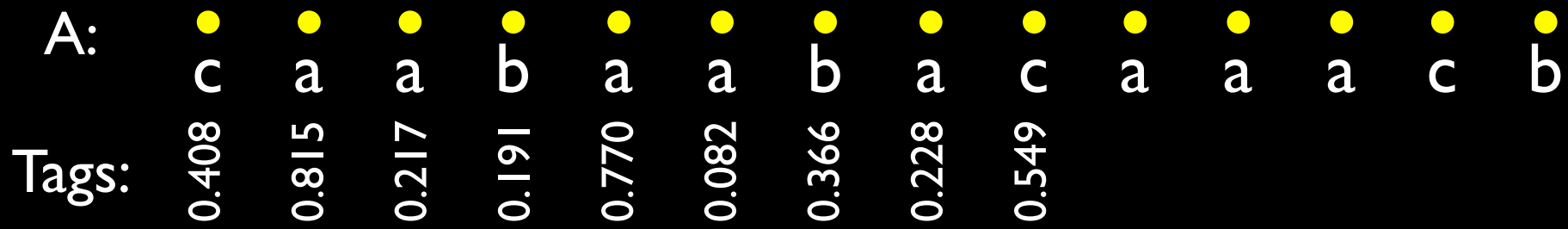
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.082, a, 2)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

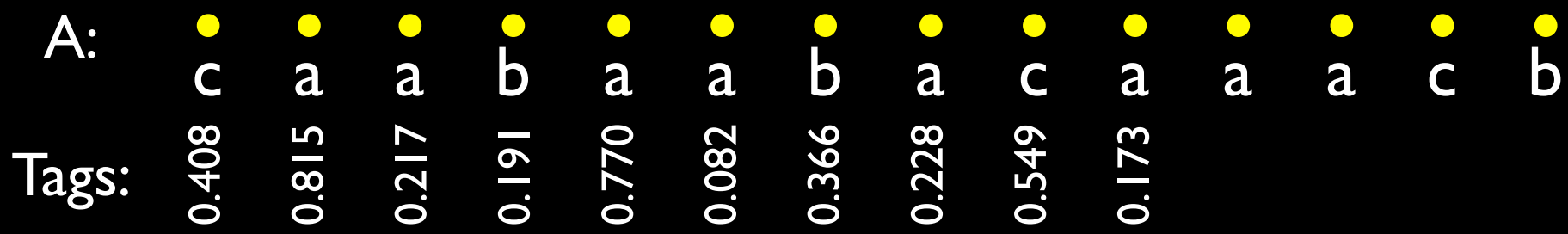
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.082, a, 2)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

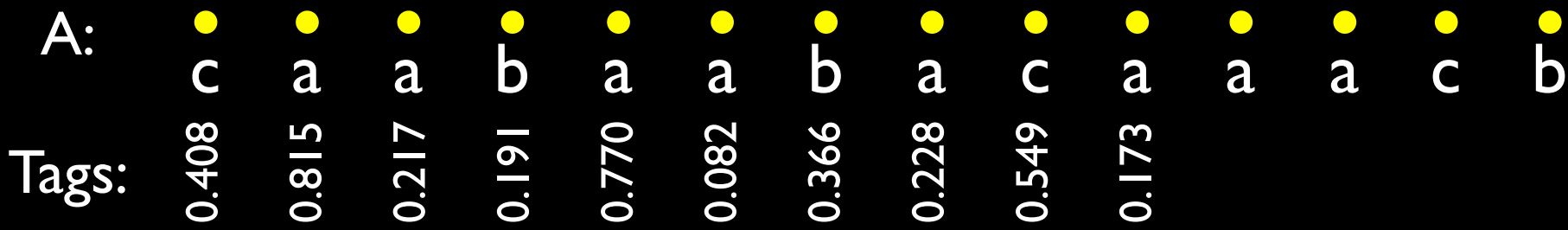
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.082, a, 2)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

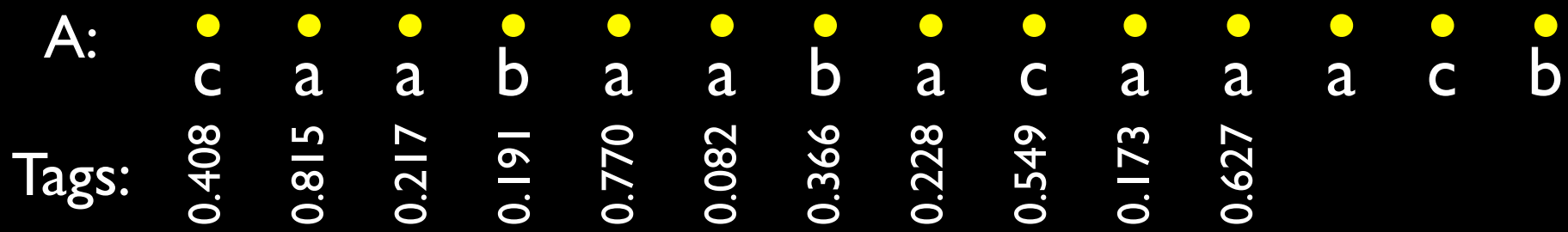
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.082, a, 3)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

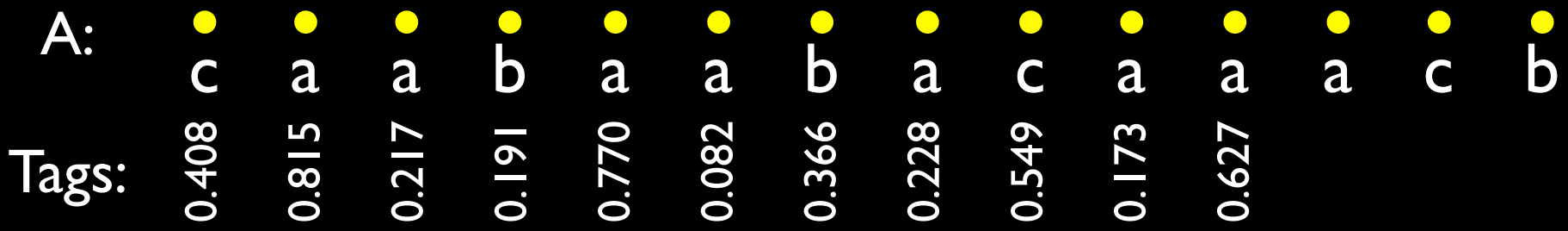
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.082, a, 3)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

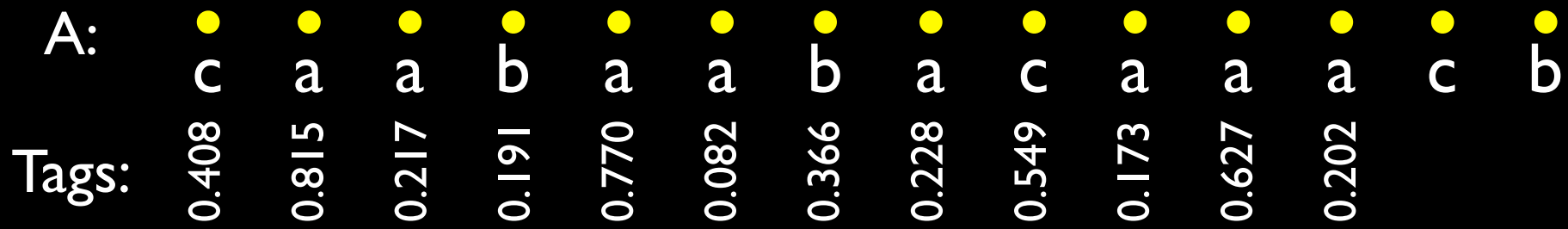
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.082, a, 4)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

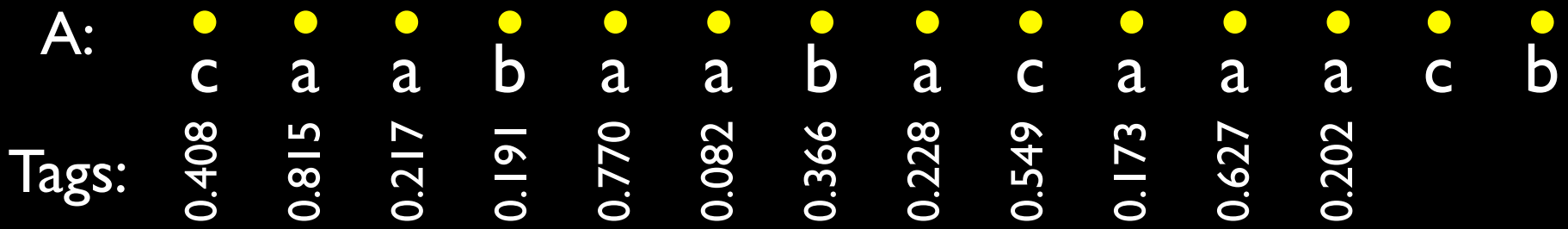
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.082, a, 4)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

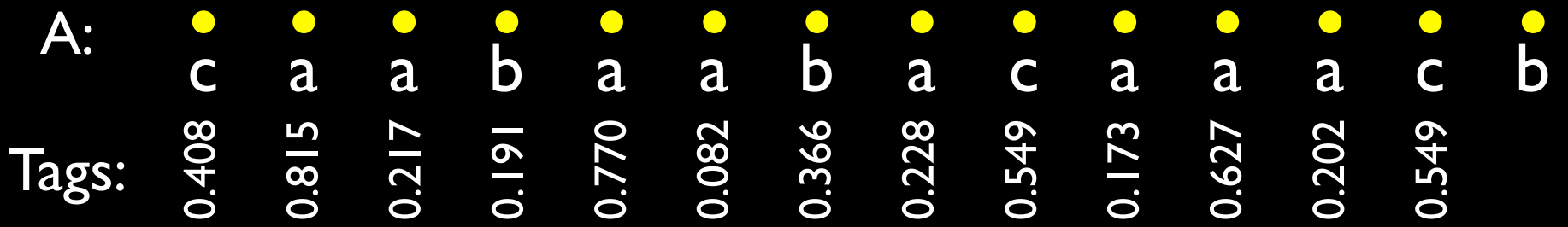
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.082, a, 5)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

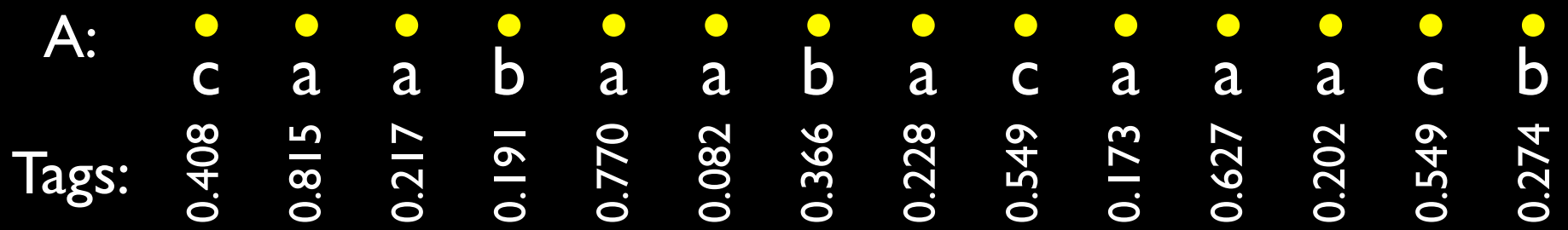
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.082, a, 5)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

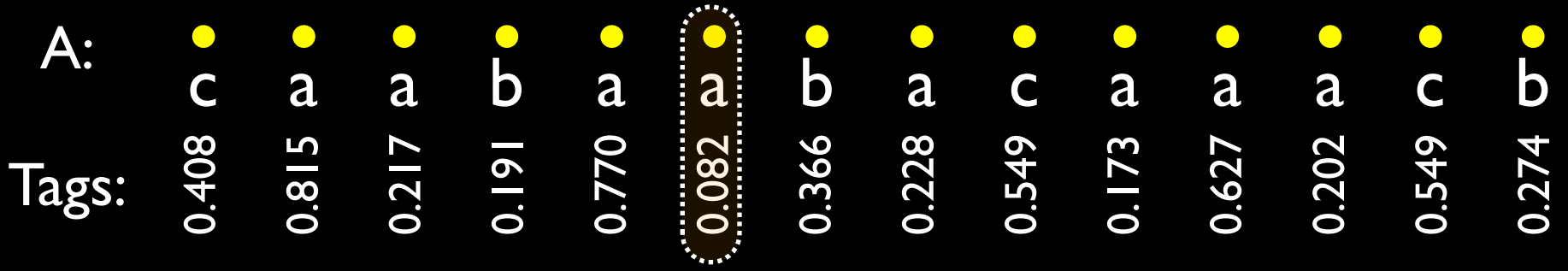
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.082, a, 5)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

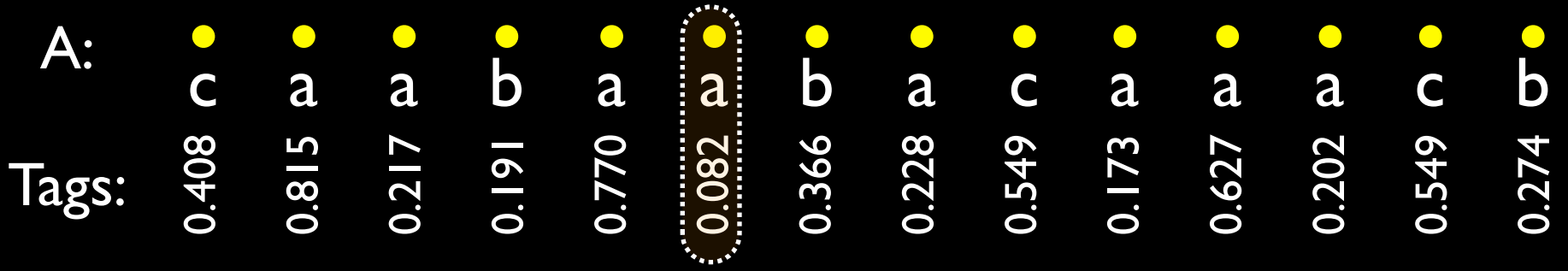
$$i \in_R A \text{ and } j \in_R [m_i]$$



(min-tag, item, repeats)= (0.082, a, 5)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

$$i \in_R A \text{ and } j \in_R [m_i]$$

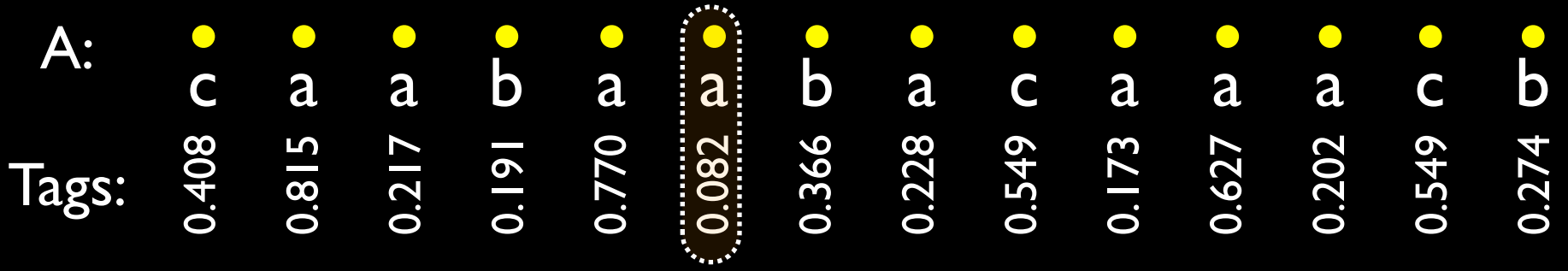


(min-tag, item, repeats)= (0.082, a, 5)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

$$i \in_R A \text{ and } j \in_R [m_i]$$

- Basic-Estimator: $X = f(j) - f(j-1)$



(min-tag, item, repeats)= (0.082, a, 5)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

$$i \in_R A \text{ and } j \in_R [m_i]$$

- Basic-Estimator: $X = f(j) - f(j-1)$
- Correct in expectation so repeat and average

$$E[X] = \sum_{i \in [m]} \frac{m_i}{m} \sum_{j \in [m_i]} \frac{1}{m_i}(f(j) - f(j-1))$$

A:	c	a	a	b	a	a	b	a	c	a	a	a	c	b
Tags:	0.408	0.815	0.217	0.191	0.770	0.082	0.366	0.228	0.549	0.173	0.627	0.202	0.549	0.274

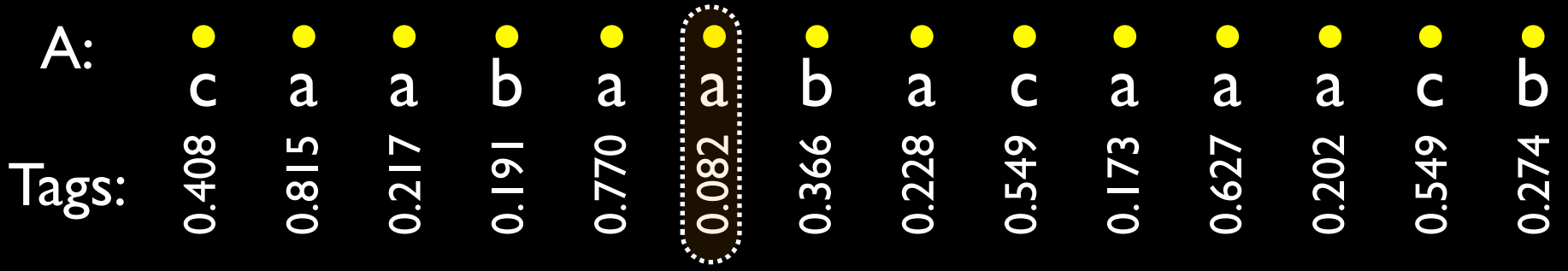
(min-tag, item, repeats)= (0.082, a, 5)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

$$i \in_R A \text{ and } j \in_R [m_i]$$

- Basic-Estimator: $X = f(j) - f(j-1)$
- Correct in expectation so repeat and average

$$3(1+\frac{a}{E[X]})^2 \epsilon^{-2} \ln(2\delta^{-1}) \frac{a+b}{a+E[X]} \text{ where } -a \leq X \leq b$$



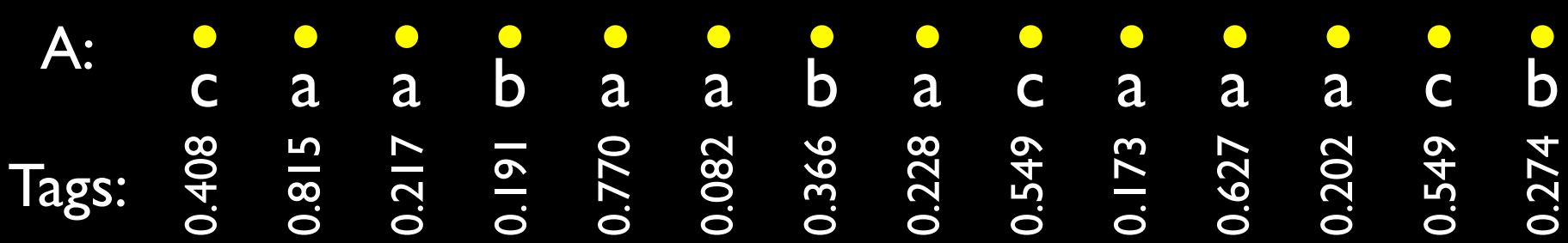
(min-tag, item, repeats)= (0.082, a, 5)

- Estimate: $\frac{1}{m}\sum f(m_i)$ where $f(x) = x \lg \frac{m}{x}$
- Define a random variable *j*:

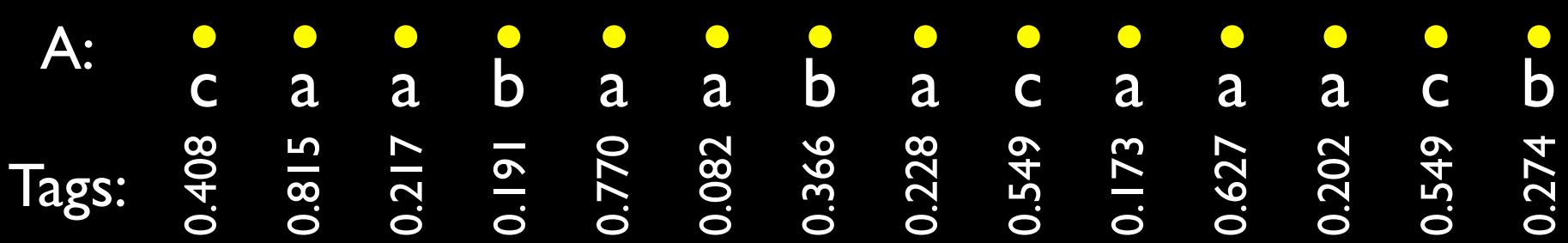
$$i \in_R A \text{ and } j \in_R [m_i]$$

- Basic-Estimator: $X = f(j) - f(j-1)$
- Correct in expectation so repeat and average

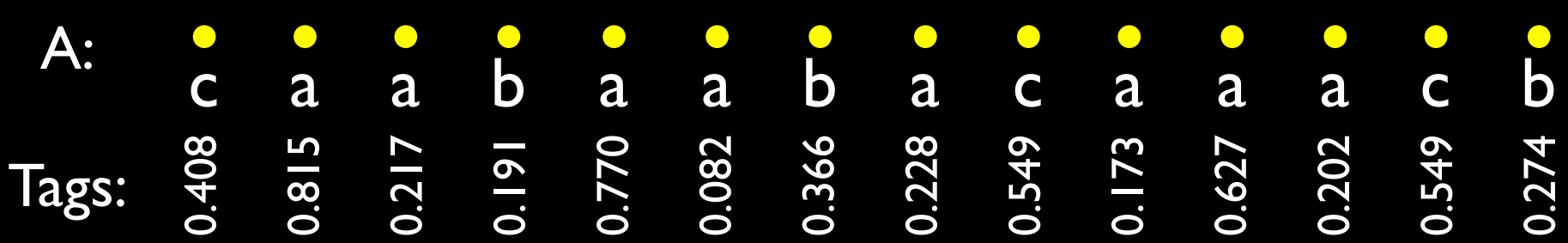
$$O((1 + H^{-1})^2 \epsilon^{-2} \log m \log \delta^{-1})$$



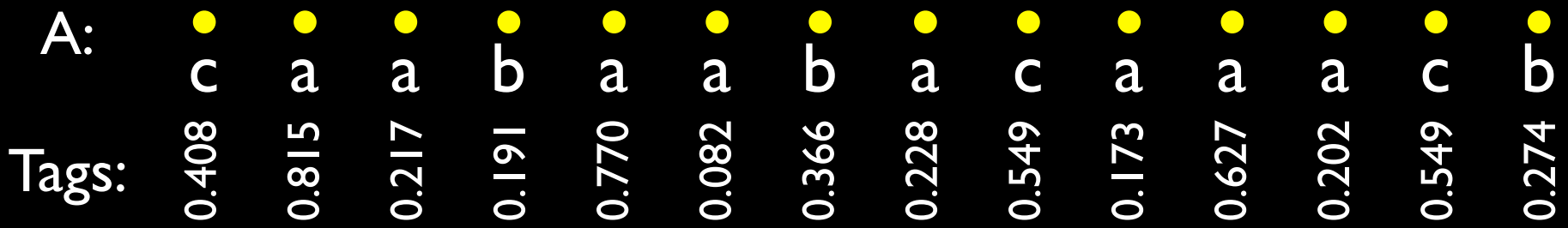
- If H is small, there exists a majority element x



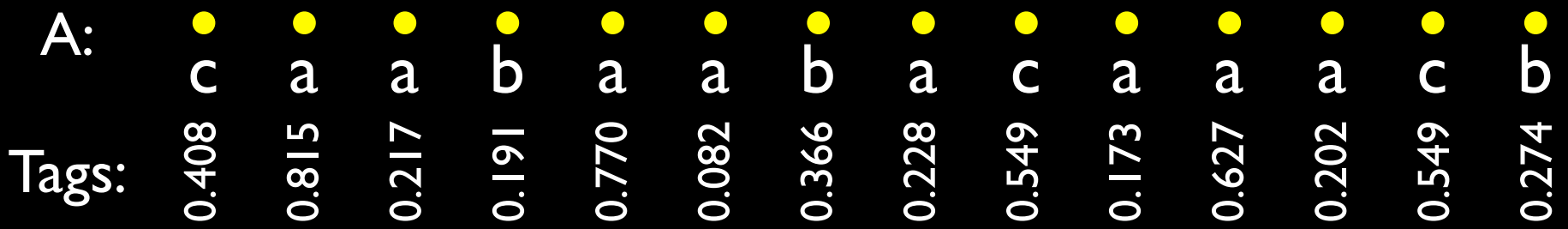
- If H is small, there exists a majority element x
- Let A′ be the sub-stream ignoring x



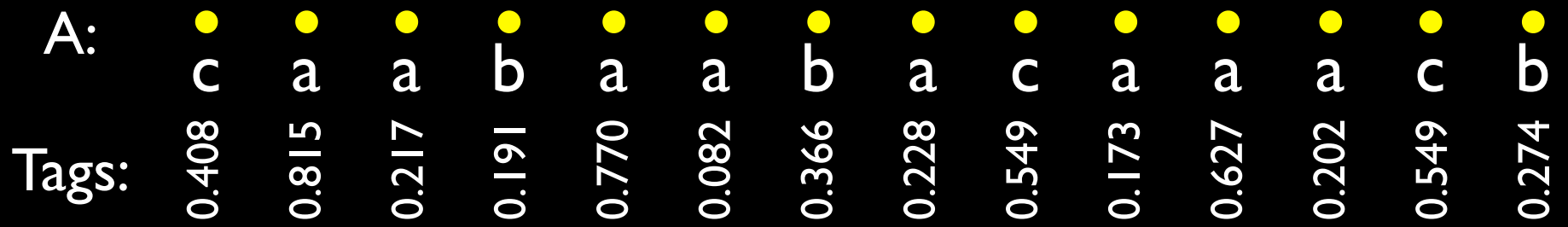
- If H is small, there exists a majority element x
- Let A′ be the sub-stream ignoring x
- Define a random variable *j*: $i \in_R A'$ and $j \in_R [m_i]$



- If H is small, there exists a majority element x
- Let A′ be the sub-stream ignoring x
- Define a random variable *j*: $i \in_R A'$ and $j \in_R [m_i]$
- Basic-Estimator: $X = f(j) - f(j-1)$

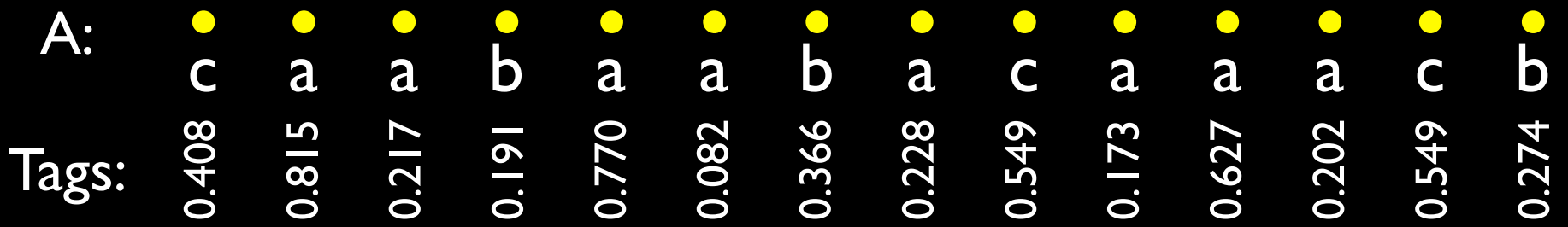


- If H is small, there exists a majority element x
- Let A′ be the sub-stream ignoring x
- Define a random variable *j*: $i \in_R A'$ and $j \in_R [m_i]$
- Basic-Estimator: $X = f(j) - f(j-1)$
- Then, $H(p) = p_x \lg p_x^{-1} + (1 - p_x)E[X]$



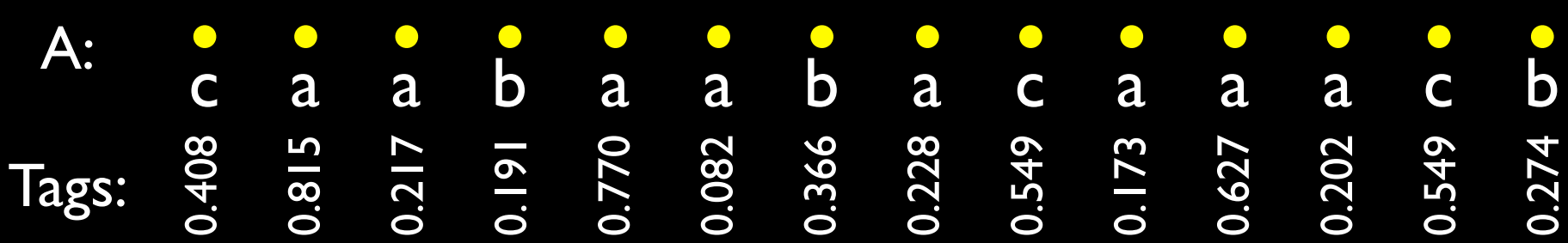
- If H is small, there exists a majority element x
- Let A′ be the sub-stream ignoring x
- Define a random variable *j*: $i \in_R A'$ and $j \in_R [m_i]$
- Basic-Estimator: $X = f(j) - f(j-1)$
- Then, $H(p) = p_x \lg p_x^{-1} + (1 - p_x) E[X]$

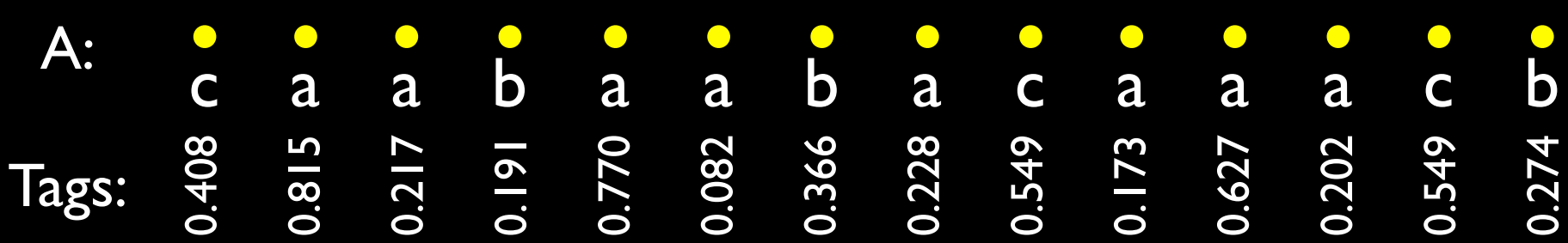
[Misra,Gries '82]



- If H is small, there exists a majority element x
- Let A′ be the sub-stream ignoring x
- Define a random variable *j*: $i \in_R A'$ and $j \in_R [m_i]$
- Basic-Estimator: $X = f(j) - f(j-1)$
- Then, $H(p) = p_x \lg p_x^{-1} + (1 - p_x) E[X]$

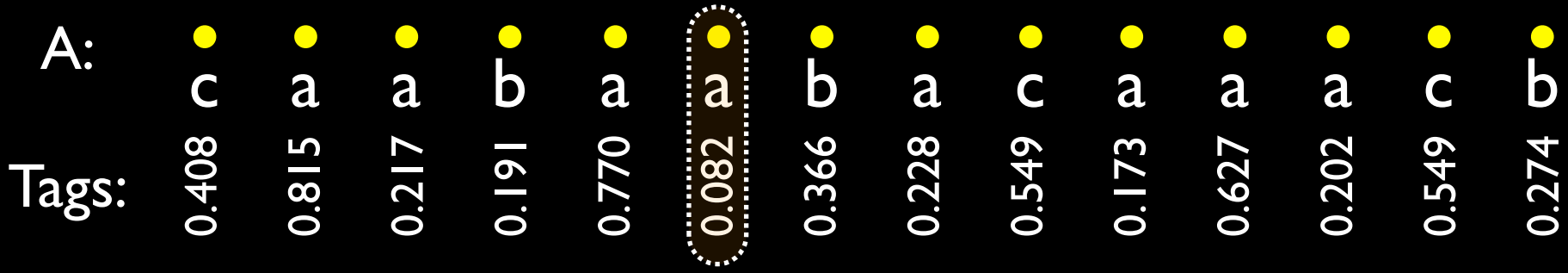
[Misra,Gries '82] Constant!


A:
c a a b a a b a c a a a c b
Tags:
0.408
0.815
0.217
0.191
0.770
0.082
0.366
0.228
0.549
0.173
0.627
0.202
0.549
0.274



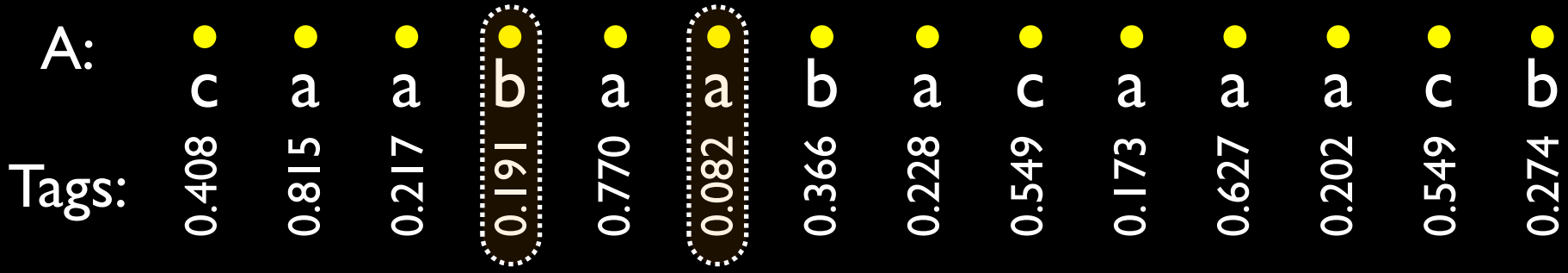
(min-tag$_1$, item$_1$, repeats$_1$)

(min-tag$_2$, item$_2$, repeats$_2$)



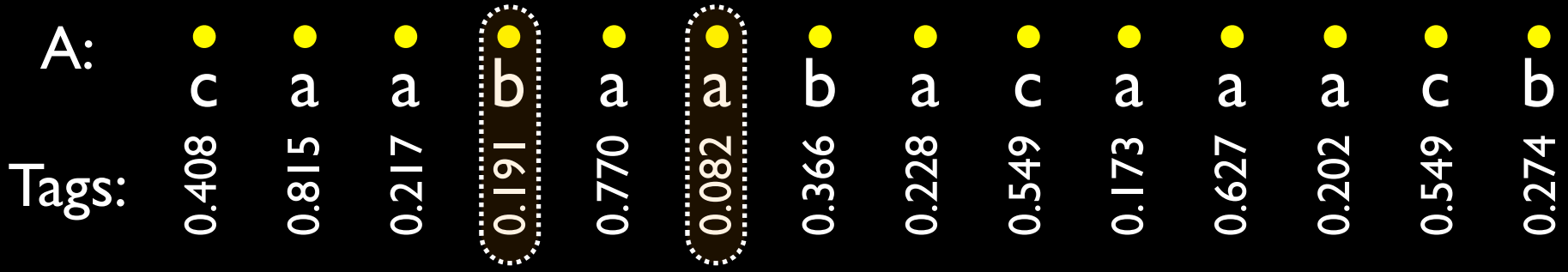
(min-tag$_1$, item$_1$, repeats$_1$)

(min-tag$_2$, item$_2$, repeats$_2$)



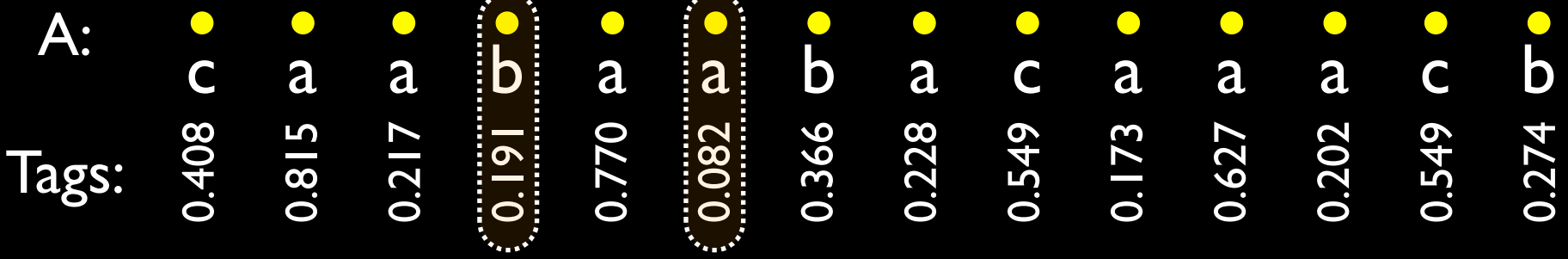
(min-tag$_1$, item$_1$, repeats$_1$)

(min-tag$_2$, item$_2$, repeats$_2$)



$(\text{min-tag}_1, \text{item}_1, \text{repeats}_1) = (0.082, a, 5)$

$(\text{min-tag}_2, \text{item}_2, \text{repeats}_2) = (0.191, b, 2)$



$(\text{min-tag}_1, \text{item}_1, \text{repeats}_1) = (0.082, a, 5)$

$(\text{min-tag}_2, \text{item}_2, \text{repeats}_2) = (0.191, b, 2)$

Update triples on seeing item y and tag t:

```
if y = item1:
    if t < min-tag1 then (min-tag1, item1, repeats1) = (t,y,1) else repeats1++
else:
    if y = item2 then repeats2++
    if t < min-tag1:
        (min-tag2, item2, repeats2) = (min-tag1, item1, repeats1)
        (min-tag1, item1, repeats1) = (t,y,1)
    else:
        if t < min-tag2 then (min-tag2, item2, repeats2) = (t,y,1)
```

Summary

Summary

- Thm: Single pass (ε,δ)-approx for entropy using $O(\varepsilon^{-2} \log m \log \delta^{-1})$ space.

Summary

- Thm: Single pass (ε,δ)-approx for entropy using $O(\varepsilon^{-2} \log m \log \delta^{-1})$ space.
- Thm: Higher-order entropy:

$$H_1 = \sum_i \frac{m_i}{m} \sum_j \frac{m_{ij}}{m_i} \lg \frac{m_i}{m_{ij}}$$

Needs $\Omega(n/\log n)$ space but can additive approx.

Summary

- Thm: Single pass (ε,δ)-approx for entropy using $O(\varepsilon^{-2} \log m \log \delta^{-1})$ space.
- Thm: Higher-order entropy:

$$H_1 = \sum_i \frac{m_i}{m} \sum_j \frac{m_{ij}}{m_i} \lg \frac{m_i}{m_{ij}}$$

  Needs $\Omega(n/\log n)$ space but can additive approx.
- Thm: Drunkards-walk entropy:

$$H_G = \frac{1}{2m} \sum_v d_v \lg d_v$$

  Single pass (ε,δ)-approx using $O(\varepsilon^{-2} \log^2 n \log^2 \delta^{-1})$ space.

1. Thesis Overview
2. Some Algorithms
3. Some Lower-Bounds

1. Thesis Overview
2. Some Algorithms
3. Some Lower-Bounds

Communication Complexity

Pointer Chasing: Multi-Value and Multi-Party

Constructing BFS Trees and Finding the Median

Communication Complexity

[Yao '78] [Nisan, Wigderson '93] [Jain, Radhakrishnan, Sen '03]

Communication Complexity

[Yao '78] [Nisan, Wigderson '93] [Jain, Radhakrishnan, Sen '03]

Alice

x

Bob

y

Communication Complexity

[Yao '78] [Nisan, Wigderson '93] [Jain, Radhakrishnan, Sen '03]

Alice

x

How many bits need to be sent for the final recipient to learn $f(x,y)$?

Bob

y

Communication Complexity

[Yao '78] [Nisan, Wigderson '93] [Jain, Radhakrishnan, Sen '03]

Alice
x

How many bits need to be sent for the final recipient to learn f(x,y)?

Bob
y

- One-round or Multi-round?
- Two-party or Multi-party?
- Randomized or Deterministic?

Alice

binary string x

Bob

binary string y

Alice

binary string x

Is Hamming distance $\Delta(x,y)$ less than $n/2$ or at least $n/2+\sqrt{n}$?

Bob

binary string y

Alice

binary string x

Is Hamming distance $\Delta(x,y)$ less than $n/2$ or at least $n/2+\sqrt{n}$?

Requires $\Omega(n)$ bits. [Indyk, Woodruff 03]

Bob

binary string y

Alice

binary string x

Is Hamming distance $\Delta(x,y)$ less than $n/2$ or at least $n/2+\sqrt{n}$?

Requires $\Omega(n)$ bits. [Indyk, Woodruff 03]

Bob

binary string y

- Assume $\exists$ an (ε,δ)-approx for entropy using S space

(a,x_1) (b,x_2) (c,x_3) (d,x_4) (e,x_5) (f,x_6) (g,x_7)

Alice

binary string x

Is Hamming distance $\Delta(x,y)$ less than $n/2$ or at least $n/2+\sqrt{n}$?

Requires $\Omega(n)$ bits. [Indyk, Woodruff 03]

Bob

binary string y

- Assume $\exists$ an (ε,δ)-approx for entropy using S space

Alice

binary string x

$(a,x_1)\ (b,x_2)\ (c,x_3)\ (d,x_4)\ (e,x_5)\ (f,x_6)\ (g,x_7)$ | $(a,y_1)\ (b,y_2)\ (c,y_3)\ (d,y_4)\ (e,y_5)\ (f,y_6)\ (g,y_7)$

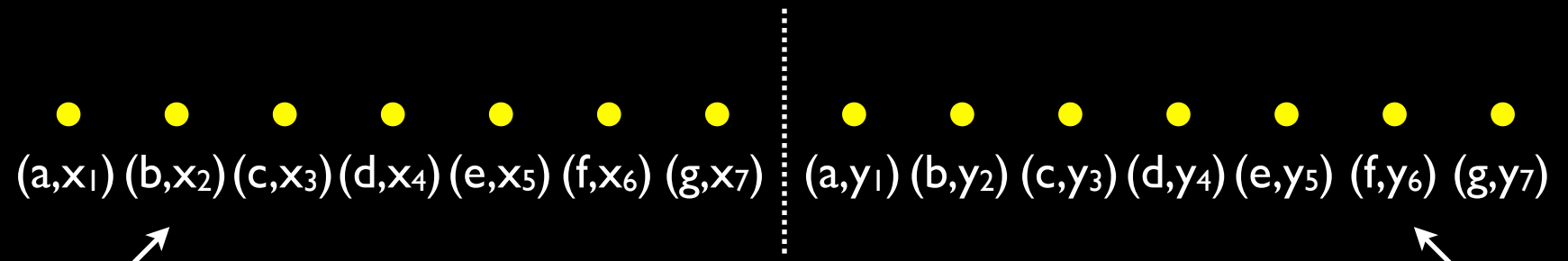
Is Hamming distance $\Delta(x,y)$ less than $n/2$ or at least $n/2+\sqrt{n}$?

Requires $\Omega(n)$ bits. [Indyk, Woodruff 03]

Bob

binary string y

- Assume $\exists$ an (ε,δ)-approx for entropy using S space



Alice

binary string x

$$H(p) = \lg n + \frac{\Delta(x,y)}{n}$$

Bob

binary string y

- Assume ∃ an (ε,δ)-approx for entropy using S space

Alice

binary string x

$$\Delta(x,y) \le n/2 \quad \Leftrightarrow \quad H(p) \le \lg n + 1/2$$
$$\Delta(x,y) \ge n/2 + \sqrt{n} \quad \Leftrightarrow \quad H(p) \ge \lg n + 1/2 + 1/\sqrt{n}$$

Bob

binary string y

- Assume ∃ an (ε,δ)-approx for entropy using S space

Alice

binary string x

$$\Delta(x,y) \le n/2 \quad \Leftrightarrow \quad H(p) \le \lg n + 1/2$$
$$\Delta(x,y) \ge n/2 + \sqrt{n} \quad \Leftrightarrow \quad H(p) \ge \lg n + 1/2 + 1/\sqrt{n}$$

Bob

binary string y

- Assume ∃ an (ε,δ)-approx for entropy using S space
- Setting $n = 3\epsilon^{-2} / \lg^2 \epsilon^{-},^1$ ratio is more than $\dfrac{1+\epsilon}{1-\epsilon}$

Alice

binary string x

$$\Delta(x,y) \le n/2 \quad \Leftrightarrow \quad H(p) \le \lg n + 1/2$$
$$\Delta(x,y) \ge n/2 + \sqrt{n} \quad \Leftrightarrow \quad H(p) \ge \lg n + 1/2 + 1/\sqrt{n}$$

Bob

binary string y

MEMORY STATE OF ALGORITHM

- Assume ∃ an (ε,δ)-approx for entropy using S space
- Setting $n = 3\epsilon^{-2} / \lg^2 \epsilon^{-}$,[1] ratio is more than $\dfrac{1+\epsilon}{1-\epsilon}$

Alice

binary string x

$$\Delta(x,y) \le n/2 \quad \Leftrightarrow \quad H(p) \le \lg n + 1/2$$
$$\Delta(x,y) \ge n/2 + \sqrt{n} \quad \Leftrightarrow \quad H(p) \ge \lg n + 1/2 + 1/\sqrt{n}$$

Bob

binary string y

(a,x_1) (b,x_2) (c,x_3) (d,x_4) (e,x_5) (f,x_6) (g,x_7) | (a,y_1) (b,y_2) (c,y_3) (d,y_4) (e,y_5) (f,y_6) (g,y_7)

MEMORY STATE OF ALGORITHM

- Assume ∃ an (ε,δ)-approx for entropy using S space
- Setting $n = 3\epsilon^{-2} / \lg^2 \epsilon^{-1}$, ratio is more than $\frac{1+\epsilon}{1-\epsilon}$
- Thm: $S=\Omega(\epsilon^{-2} / \log^2 \epsilon^{-1})$

Pointer-Chasing

Alice

function

$f_A: [m] \to [m]$

Bob

function

$f_B: [m] \to [m]$

Pointer-Chasing

Alice

function

$f_A: [m] \to [m]$

Compute

$f_A(f_B(\ldots (f_A(f_B(f_A(1)))) \ldots))$

k

Bob

function

$f_B: [m] \to [m]$

Pointer-Chasing

Alice

function
$f_A: [m] \to [m]$

Compute

$f_A(f_B(\ldots (f_A(f_B(f_A(1)))) \ldots))$

$\longleftrightarrow$ k

If Bob speaks first:

k messages: $O(k \log m)$ bits.

k-1 messages: $\Omega(m)$ bits.

[Nisan, Wigderson '93]

Bob

function
$f_B: [m] \to [m]$

Multi-Value Pointer-Chasing

Alice

d-valued function

$f_A: [m] \to [m]$

Compute

$$f_A(f_B(\ldots (f_A(f_B(f_A(1)))) \ldots))$$

$\longleftrightarrow$ k

If Bob speaks first:

k messages: $O(k\, d^k \log m)$ bits.

k-1 messages: $\Omega(d\, m)$ bits.

Bob

d-valued function

$f_B: [m] \to [m]$

Multi-Value Pointer-Chasing

Alice

d-valued function

$f_A: [m] \to [m]$

Compute

$f_A(f_B(\ldots (f_A(f_B(f_A(1)))) \ldots))$

$\longleftrightarrow$ k

If Bob speaks first:

k messages: $O(k\, d^k \log m)$ bits.

k-1 messages: $\Omega(d\, m)$ bits.

Bob

d-valued function

$f_B: [m] \to [m]$

- Proof: Relate to solving d copies of single-valued problem

Multi-Value Pointer-Chasing

Alice

d-valued function

$f_A: [m] \to [m]$

Compute

$$f_A(f_B(\ \ldots\ (f_A(f_B(f_A(1))))\ \ldots\ ))$$

k

If Bob speaks first:

k messages: $O(k\, d^k \log m)$ bits.

k-1 messages: $\Omega(d\, m)$ bits.

Bob

d-valued function

$f_B: [m] \to [m]$

- Proof: Relate to solving d copies of single-valued problem
  Appeal to "Direct-sum" Theorem [Jain, Radhakrishnan, Sen 03]

Multi-Value Pointer-Chasing

Alice

d-valued function

$f_A: [m] \to [m]$

Compute

$f_A(f_B(\ldots (f_A(f_B(f_A(1)))) \ldots))$

$\longleftrightarrow$ k

If Bob speaks first:

k messages: $O(k\, d^k \log m)$ bits.

k-1 messages: $\Omega(d\, m)$ bits.

Bob

d-valued function

$f_B: [m] \to [m]$

- Proof: Relate to solving d copies of single-valued problem
  Appeal to "Direct-sum" Theorem [Jain, Radhakrishnan, Sen 03]
- Can be reduced to the problem of finding BFS trees

Multi-Party Pointer-Chasing

Alice
function
$f_A: [m] \to [m]$

Bob
function
$f_B: [m] \to [m]$

Claude
function
$f_C: [m] \to [m]$

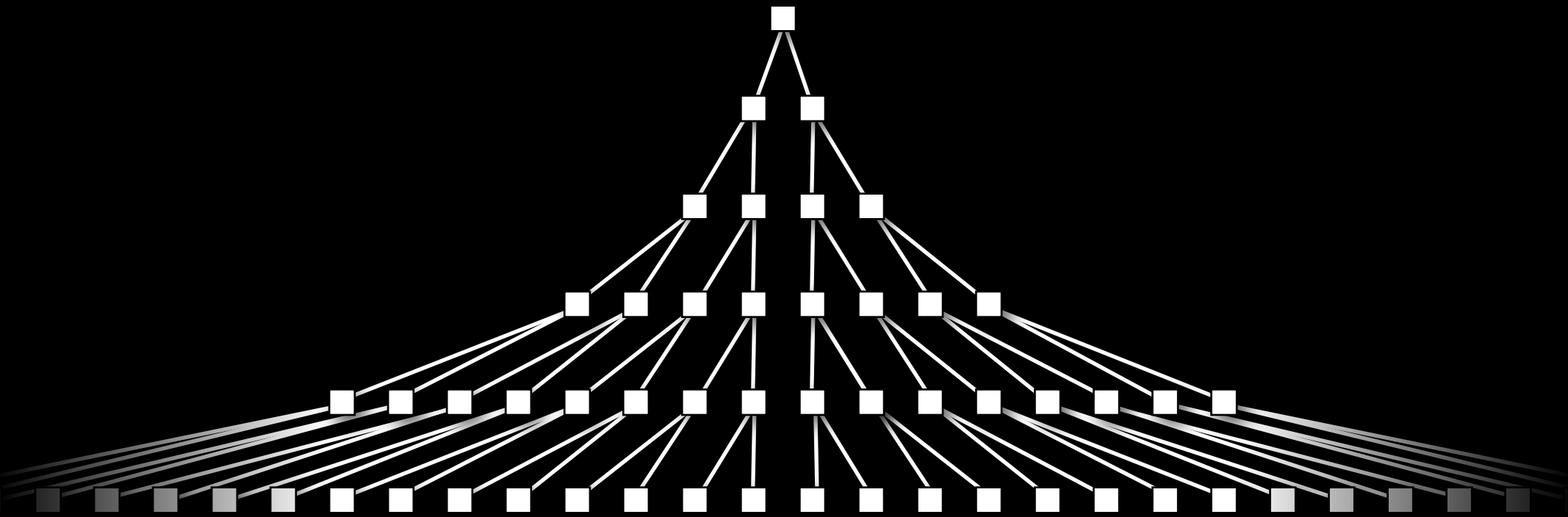
...

Zebedee
function
$f_Z: [m] \to [m]$

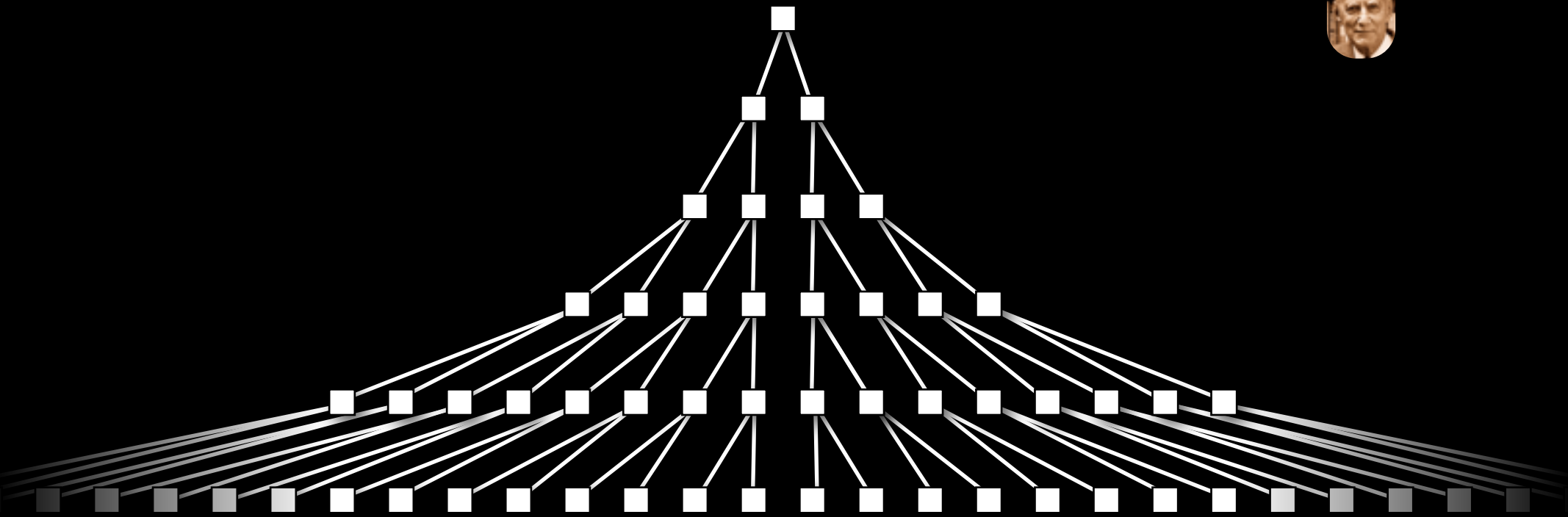
Compute $f_Z(\ \ldots f_C(f_B(f_A(1))) \ldots)$

Order of speaking "Z, ..., C, B, A"
k rounds: $O(k \log m)$ bits.
k-1 rounds: $\Omega(m)$ bits.

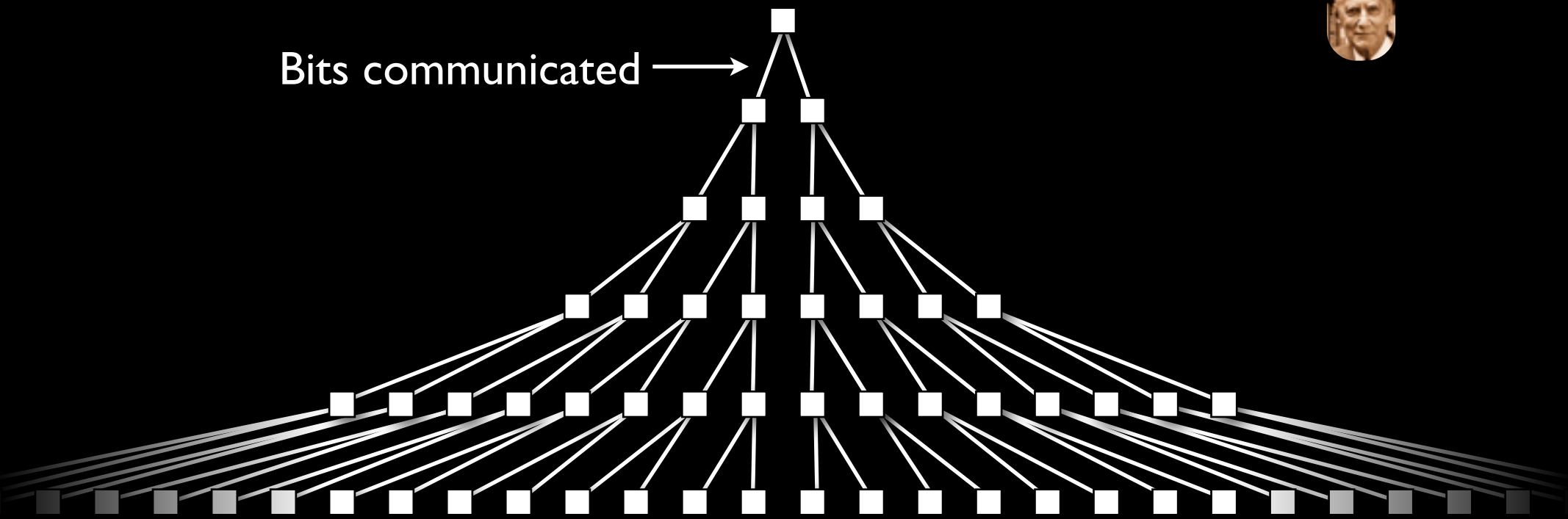
Proof Sketch



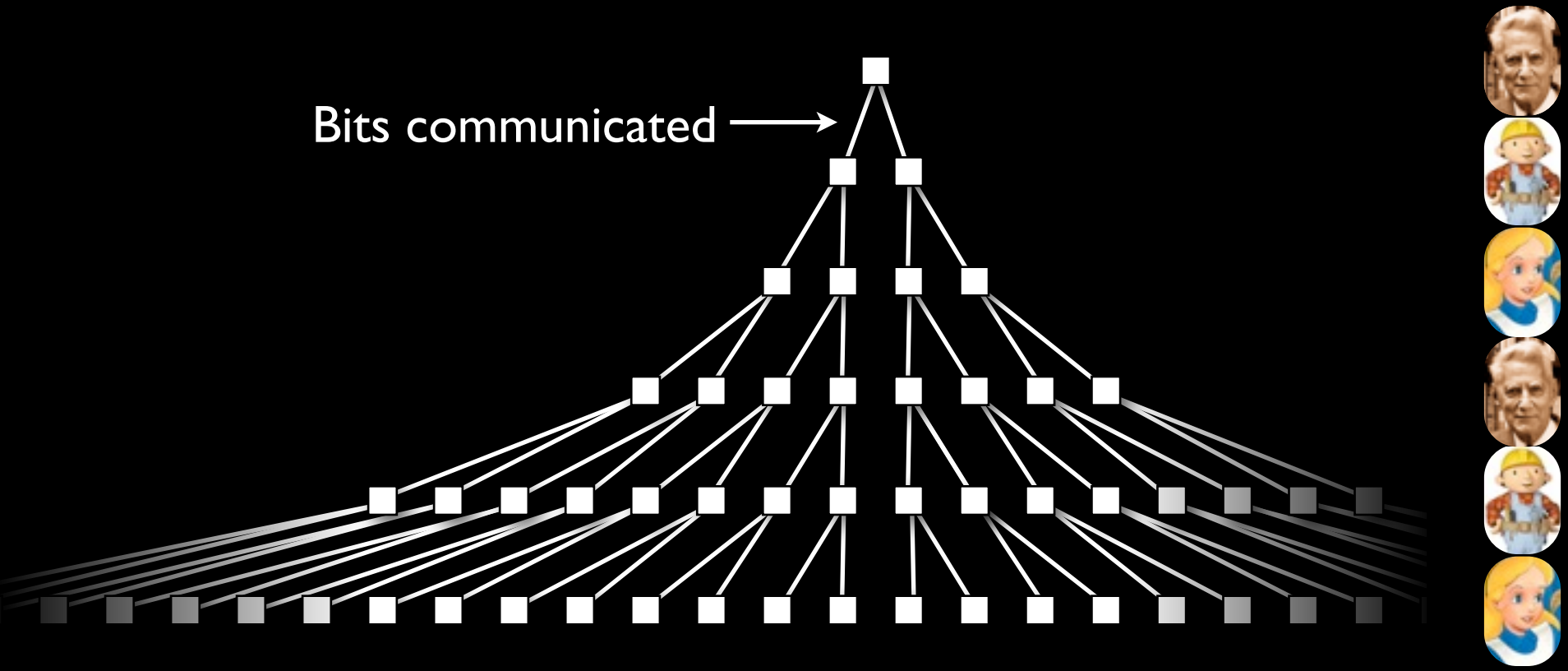
Proof Sketch



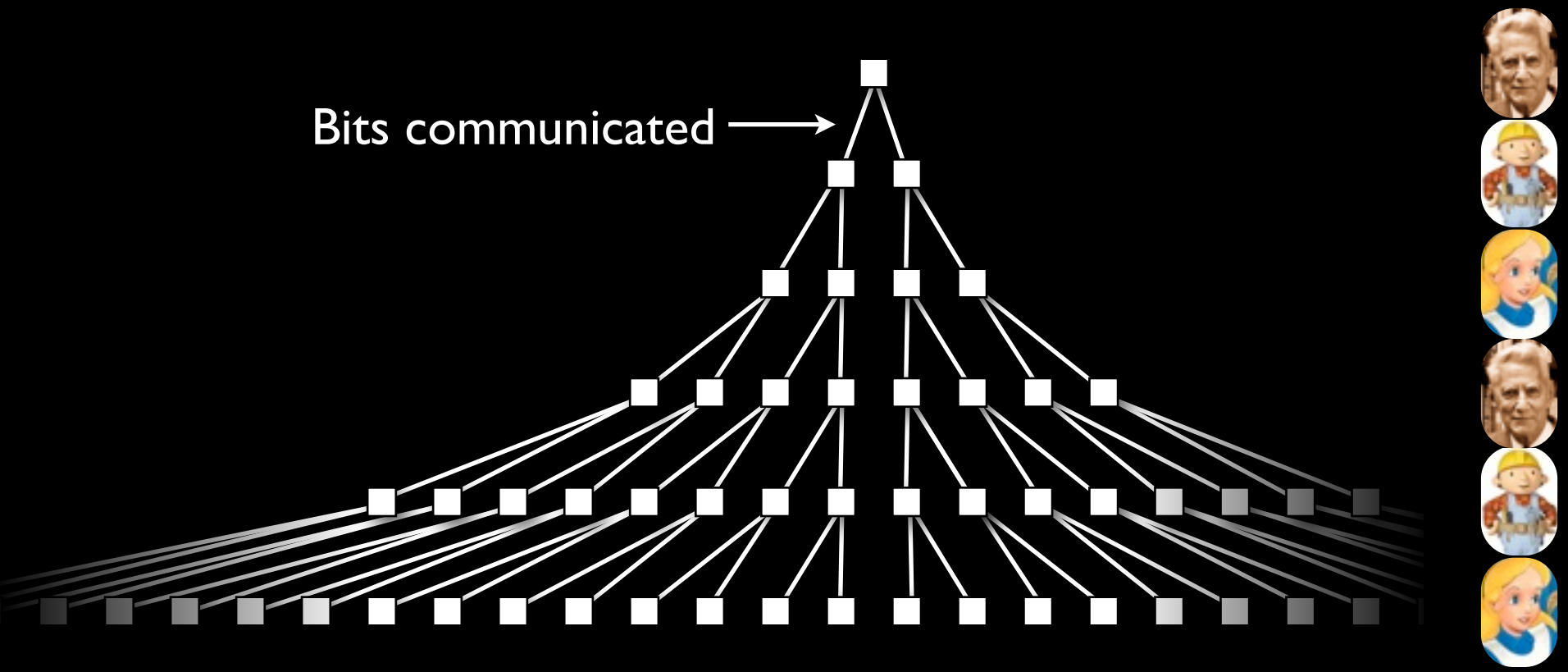
Proof Sketch



Proof Sketch

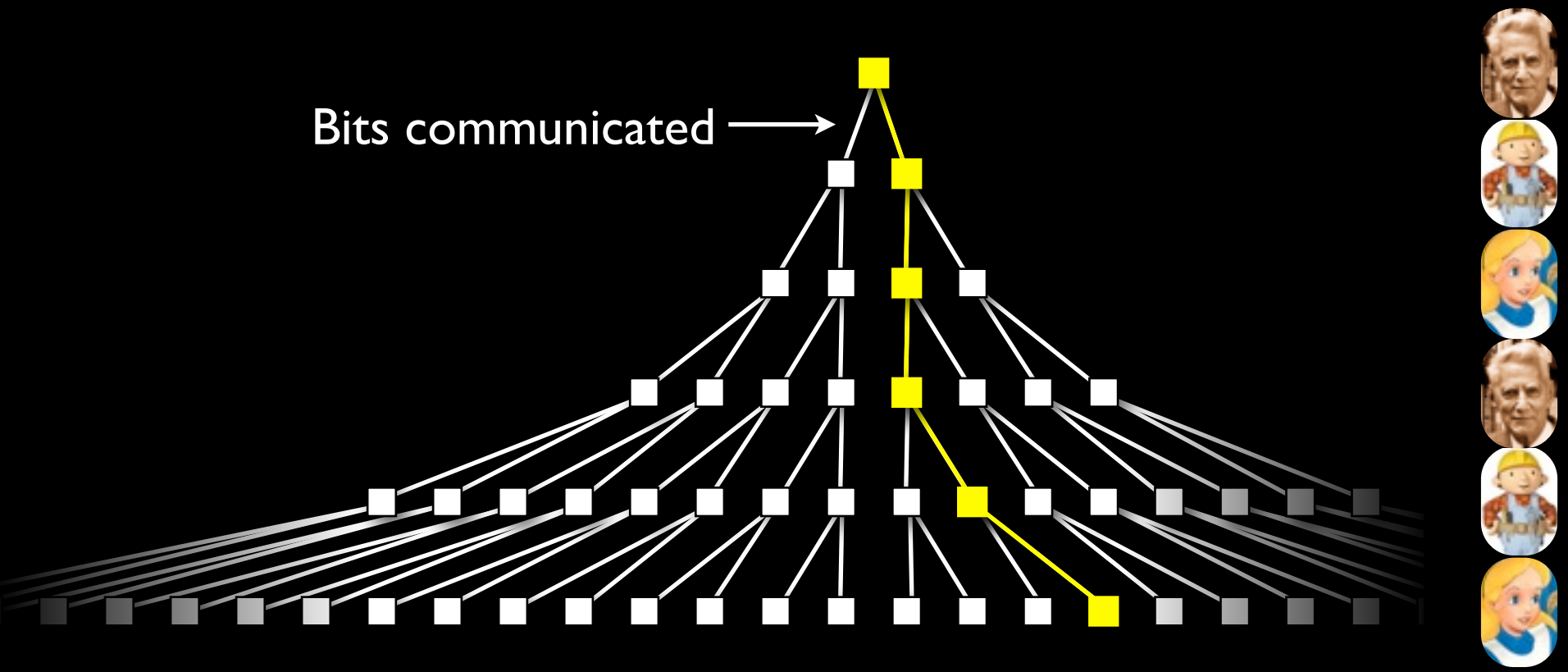


Proof Sketch



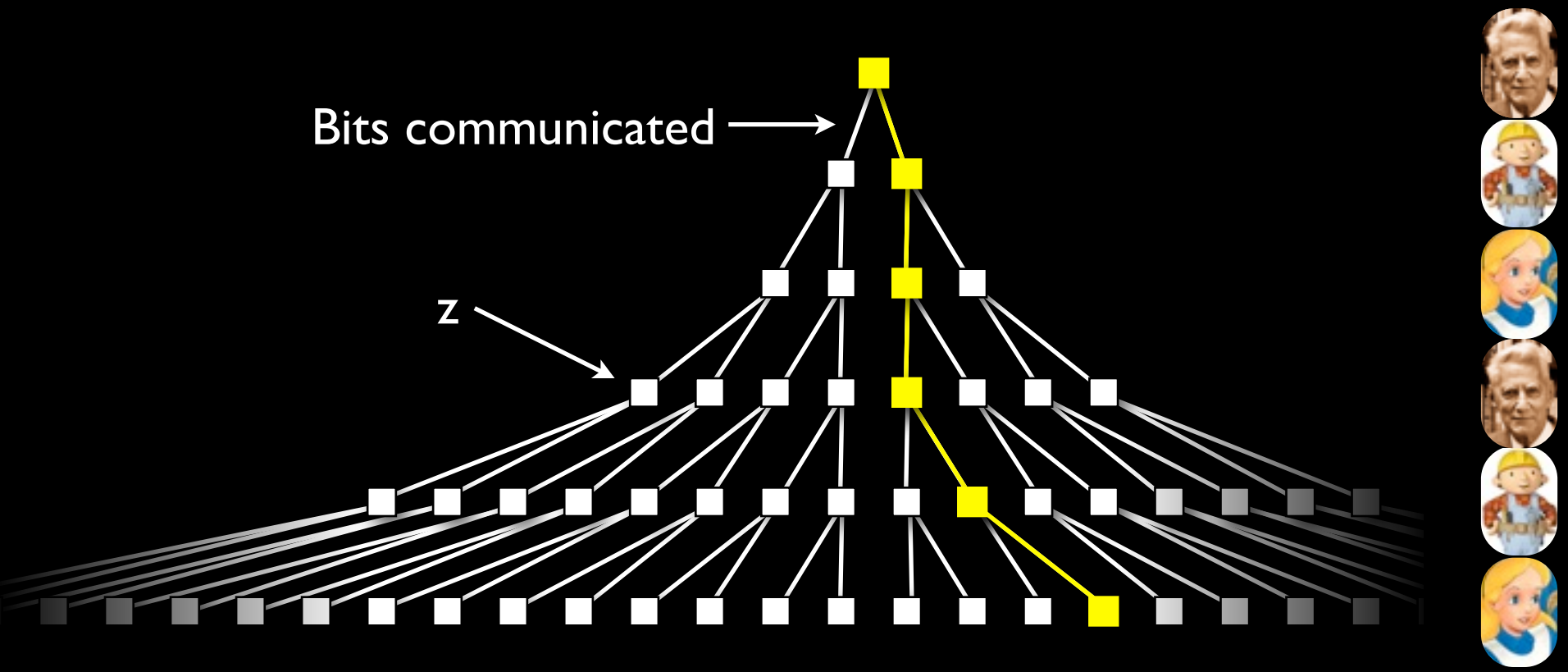
- Consider deterministic protocols and random f_A, f_B, f_C

Proof Sketch



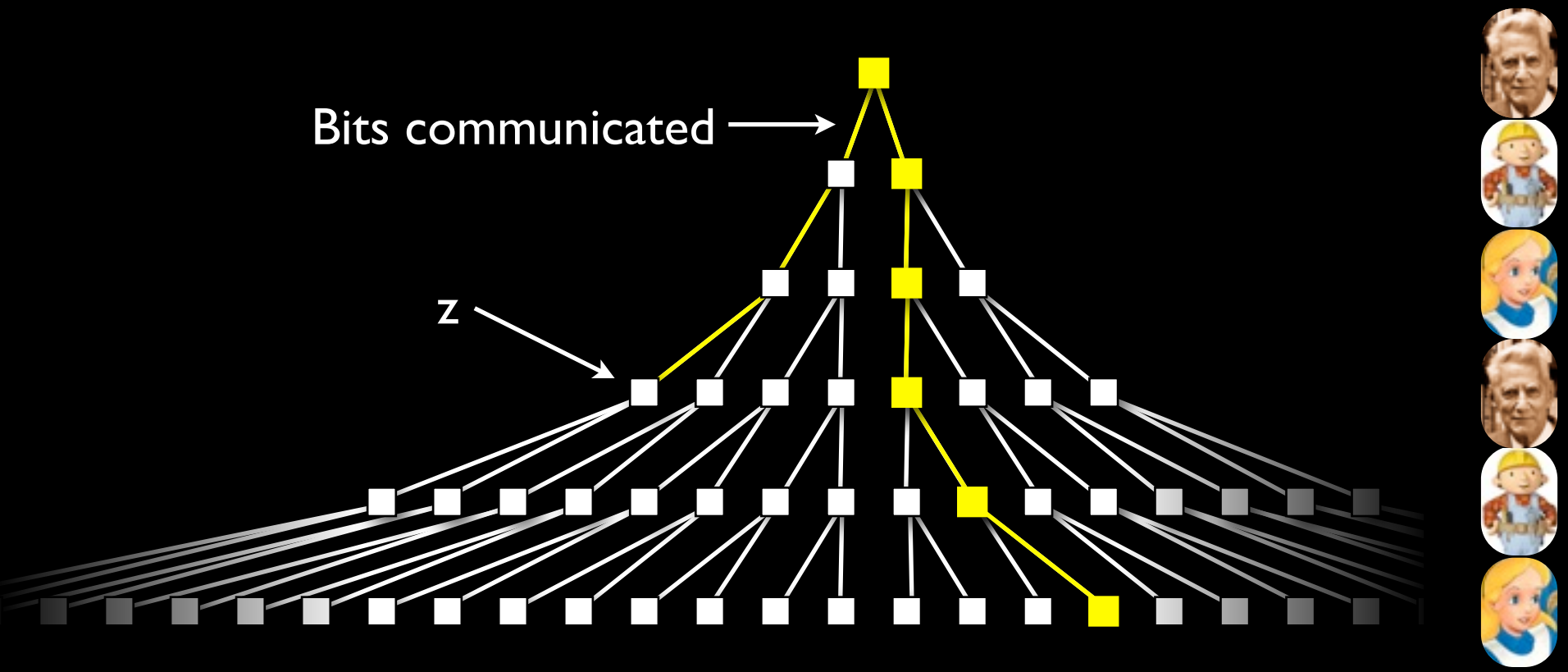
- Consider deterministic protocols and random f_A, f_B, f_C

Proof Sketch



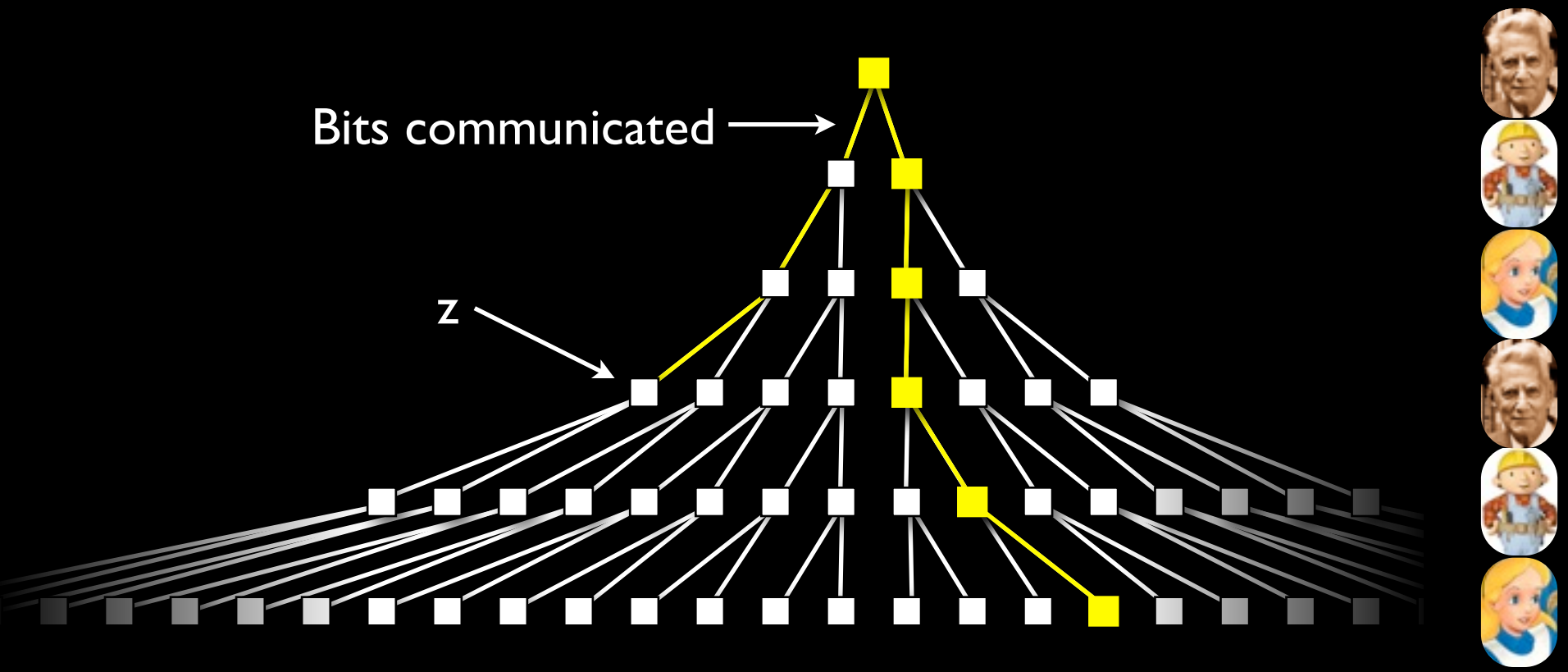
- Consider deterministic protocols and random f_A, f_B, f_C
- At each node z defines random variables $z(f_A)$, $z(f_B)$, $z(f_C)$

Proof Sketch



- Consider deterministic protocols and random f_A, f_B, f_C
- At each node z defines random variables $z(f_A)$, $z(f_B)$, $z(f_C)$

Proof Sketch



- Consider deterministic protocols and random f_A, f_B, f_C
- At each node z defines random variables $z(f_A)$, $z(f_B)$, $z(f_C)$
- Induction: For each z, entropy of variables is high.

Reduction to Selection

f_A: [m]→[m]

f_B: [m]→[m]

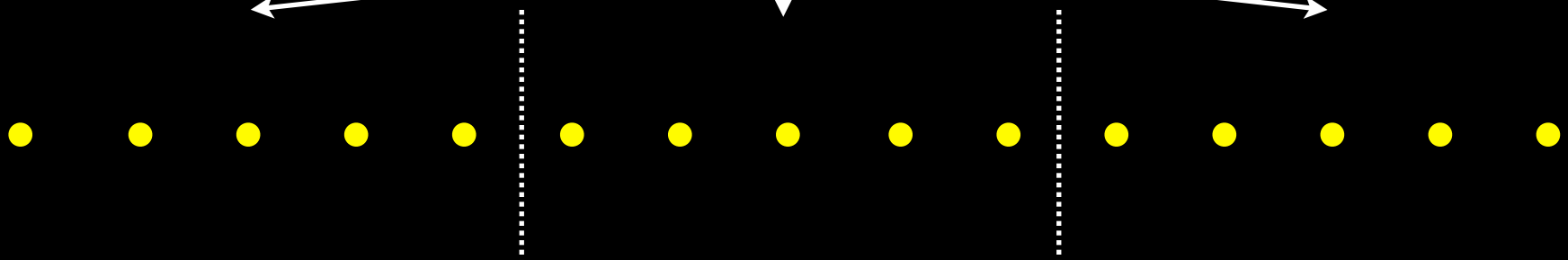
f_C: [m]→[m]

Reduction to Selection

f_A: [m]→[m]

f_B: [m]→[m]

f_C: [m]→[m]

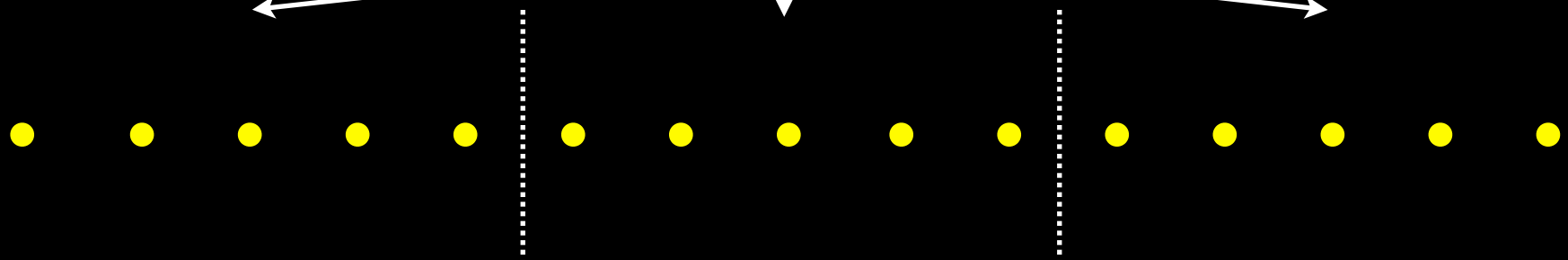


Reduction to Selection

f_A: [m]→[m]

f_B: [m]→[m]

f_C: [m]→[m]



Median=$f_A(1)$ $f_B(f_A(1)$ $f_C(f_B(f_A(1)))$

Reduction to Selection

0 0 0 x (3-$f_A(1)$) x 5

1 0 0 x (3-$f_B(1)$)

1 1 $f_C(1)$
1 2 $f_C(2)$
1 3 $f_C(3)$

1 4 0 x ($f_B(1)$-1)

2 0 0 x (3-$f_B(2)$)

2 1 $f_C(1)$
2 2 $f_C(2)$
2 3 $f_C(3)$

2 4 0 x ($f_B(2)$-1)

3 0 0 x (3-$f_B(3)$)

3 1 $f_C(1)$
3 2 $f_C(2)$
3 3 $f_C(3)$

3 4 0 x ($f_B(3)$-1)

4 0 0 x ($f_A(1)$-1) x 5

VALUE

Summary

- Thm: Any single pass $(\varepsilon, 1/4)$-approx for entropy requires $\Omega(\varepsilon^{-2} / \log^2 \varepsilon^{-1})$ space.
- Thm: Computing first t layers of a BFS-tree requires either $t/2$ passes or $\Omega(n^{1+1/t})$ space.
- Thm: Finding the median requires either p passes or $\Omega(n^{1/p})$ space.

Future Directions

Future Directions

- Little things: Factors in space and passes

Future Directions

- Little things: Factors in space and passes
- Medium things: Random-order BFS trees,
  Lower-bounds for random-order streams,
  Fully characterize the "sketch-able" distances.

Future Directions

- Little things: Factors in space and passes
- Medium things: Random-order BFS trees,
  Lower-bounds for random-order streams,
  Fully characterize the "sketch-able" distances.
- Big thing: Space-Efficient Sampling!

Thanks

Thanks

Advisor and Mentors:

Sampath, Sudipto,
Sasha, & Bruce

Thanks

Advisor and Mentors:

Sampath, Sudipto,
Sasha, & Bruce

Committee:

Sudipto, Piotr,
Michael, & Sanjeev

Thanks

Advisor and Mentors:

Sampath, Sudipto,
Sasha, & Bruce

Committee:

Sudipto, Piotr,
Michael, & Sanjeev

Theory Group (Past & Present):

Stan, Milan, Yael, Boulos, Niel,
Kuku, Sid, & Mirko

Thanks

Advisor and Mentors:

Sampath, Sudipto, Sasha, & Bruce

Committee:

Sudipto, Piotr, Michael, & Sanjeev

Theory Group (Past & Present):

Stan, Milan, Yael, Boulos, Niel, Kuku, Sid, & Mirko

Co-Authors:

Deepak, Stan, Sasha, Tugkan, Amit, Graham, Joan, Peter, Boulos, Piotr, Sampath, Sanjeev, Keshav, Eduardo, Jeff, Bruce, Sid, Suresh, Jian, & Zhengyuan

Thanks

Advisor and Mentors:

Sampath, Sudipto, Sasha, & Bruce

Committee:

Sudipto, Piotr, Michael, & Sanjeev

Theory Group (Past & Present):

Stan, Milan, Yael, Boulos, Niel, Kuku, Sid, & Mirko

Co-Authors:

Deepak, Stan, Sasha, Tugkan, Amit, Graham, Joan, Peter, Boulos, Piotr, Sampath, Sanjeev, Keshav, Eduardo, Jeff, Bruce, Sid, Suresh, Jian, & Zhengyuan

"Normal" People:

Friends, family, visitors, & especially Christie

Questions?