Overview and basic techniques

Luca Becchetti

“Sapienza” Università di Roma – Rome, Italy

June 18, 2011
1. Course overview
2. Massive data sets
3. Probability basics
4. Expectation and variance of discrete variables
5. Concentration of measure
6. Dictionaries and hashing
Goals

- Course only provides an overview of a few areas
- Understand problems in handling massive data sets
- Understand basic principles in addressing these issues
- Perform a deeper study of an area of choice among eligible ones
  - Understand problems
  - Understand basic techniques
  - Understand key results

Exam (2008/2009)

- Written exam
- Answer 2 out of a collection of 10 possible published questions (7.5 points each)
- Answer a few questions about a research paper (25 points)
- Example: explain reference scenario/key results/techniques used/...
Contents and expected preparation

Topics

1. Basic techniques and tools
   - Basic probabilistic tools
   - Brief review of hashing

2. Bloom filter - A compact database summary
   - Properties and applications

3. Data streaming
   - Applications and computational model
   - Some key results

Your expected preparation

- Good understanding of 1
- Fair understanding of all topics covered in the course
  - Lessons + review of main references
- In-depth knowledge of one topic of choice
  - Main references + teacher’s suggested readings
More about the exam

The course

- Elective Course in Computer Networks consists of 3 CFU units
  - CFU: Credito Formativo Universitario
- Students who attend the course may pass the exam for 1 to 4 units
- An exam has to be passed for each chosen unit

Mark

- A mark from 18 to 30 in each unit
- Final mark is average of votes achieved in all chosen units
- Marks received in single units are communicated to the responsible person, Prof. Marchetti-Spaccamela
Evaluation criteria

- Quality of presentation
  - How you present the topic, the language used etc.
  - The organization of your presentation
  - How clear and rigorous is your presentation
  - Adequacy of references

- Your understanding of the topic
  - How confident you are with the topic
  - How able you are to discuss your topic critically, to answer questions, to address related topics
  - How well you understand the basic underlying principles
  - Your ability to outline potential or motivating application scenarios behind the topic considered
<table>
<thead>
<tr>
<th>Year</th>
<th>TB/month</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>1.0</td>
</tr>
<tr>
<td>1991</td>
<td>2.0</td>
</tr>
<tr>
<td>1992</td>
<td>4.4</td>
</tr>
<tr>
<td>1993</td>
<td>8.3</td>
</tr>
<tr>
<td>1994</td>
<td>16.3</td>
</tr>
<tr>
<td>1995</td>
<td>?</td>
</tr>
<tr>
<td>1996</td>
<td>1,500</td>
</tr>
<tr>
<td>1997</td>
<td>2,500 - 4,000</td>
</tr>
<tr>
<td>1998</td>
<td>5,000 - 8,000</td>
</tr>
<tr>
<td>1999</td>
<td>10,000 - 16,000</td>
</tr>
<tr>
<td>2000</td>
<td>20,000 - 35,000</td>
</tr>
<tr>
<td>2001</td>
<td>40,000 - 70,000</td>
</tr>
<tr>
<td>2002</td>
<td>80,000 - 140,000</td>
</tr>
</tbody>
</table>

- Traffic explosion in past years [Muthukrishnan, 2005]
  - 30 billions emails, 1 billions SMS, IMs daily (2005)
  - \( \approx 1 \text{ billion packets/router x hr} \)
Logs

- **SNMP**: (Router ID, Interface ID, Timestamp, Bytes sent since last obs.)

- **Flow**: (Source IP, Dest IP, Start Time, Duration, No. Packets, No. Bytes)

- (Source IP, Dest IP, Src/Dest Port Numbers, Time, No. Bytes)
Motivations/3

Database access

- Huge amounts of data
- Large number of retrieve requests per sec.
- DB index in main memory
- May be too large to fit in or for fast access
Challenges [Muthukrishnan, 2005]

- 1 link with 2 Gb/s. Say avg packet size is 50 bytes
- Number of pkts/sec = 5 Million
- Time per pkt = 0.2 µsec (time available for processing)
- If we capture pkt headers per packet: src/dest IP, time, no of bytes, etc. at least 10 bytes
- Space per second is 50 MB. Space per day is 4.5 TB per link
- ISPs have hundreds of links.

Focus is on solutions for real applications

Note: we seek solutions that work in practice → easy to implement, require small space, allow fast updates and queries
Events and probability

- Sample space $O$
- Event: subset $E \subseteq O$ of outcomes that satisfy given condition
- In the example: choose a ball uniformly at random
  - $E = \text{(A yellow ball is picked)}$

Example: bin with balls extracted at random

$$\Pr[E] = 1/4$$
Axioms of probability

- $\mathcal{F}$ is the set of possible events
  - Example of event: A yellow or a green ball is extracted $\rightarrow$ subset of yellow and green balls

Probability function: any function $\mathbf{P} : \mathcal{F} \rightarrow \mathbb{R}$

Axioms of probability:

- For every $E \in \mathcal{F}$: $0 \leq \mathbf{P}[E] \leq 1$
- $\mathbf{P}[O] = 1$
- For any set $E_1 \ldots , E_n$ of mutually disjoint events ($E_i \cap E_h = \emptyset$, $\forall i, h$): $\mathbf{P}[\bigcup_{i=1}^{n} E_i] = \sum_{i=1}^{n} \mathbf{P}[E_i]$
Some basic facts

For any two events $E_1, E_2$:

- $\mathbb{P}[E_1 \cup E_2] = \mathbb{P}[E_1] + \mathbb{P}[E_2] - \mathbb{P}[E_1 \cap E_2]$
- Formula above generalizes

In general:

**Fact**

$$\mathbb{P}[\bigcup_{i=1}^{n} E_i] \leq \sum_{i=1}^{n} \mathbb{P}[E_i]$$

**Conditional probability**

Conditional probability that $E$ occurs given that $F$ occurs:

$$\mathbb{P}[E \mid F] = \frac{\mathbb{P}[E \cap F]}{\mathbb{P}[F]}$$

Events $E_1, \ldots, E_k$ are mutually independent if and only if, for every $I \subseteq \{1, \ldots, k\}$:

$$\mathbb{P}[\bigcap_{i \in I} E_i] = \prod_{i \in I} \mathbb{P}[E_i]$$

For two events $E, F$ this implies: $\mathbb{P}[E \mid F] = \mathbb{P}[E]$
Warm up questions

**Q1:** Consider a bin with an equal number $n/2$ of white and black balls. Assume $w$ white and $b$ black balls have been extracted with replacement.

- What is the probability that the next ball extracted is white?
- What does the sample space look like?

**Q2:** Answer again the first question if extraction occurs without replacement.

**Q3:** $n$ bits are transmitted in sequence over a line on which every bit has probability $1/2$ of being flipped due to noise, independently of all other bits in the sequence. For $k > 0$, give an upper bound on the probability that there is a sequence of at least $\log_2 n + k$ consecutive inversions (see also exercise 1.11 in [Mitzenmacher and Upfal, 2005])
Q3: sketch of solution

- Let $X_i = 1$ if $i$-th bit flipped, 0 otherwise
- Let $E_i = (\land_{t=i}^{i+\log_2 n+k} X_i = 1)$

Solution

\[
P[\text{At least } \log_2 n + k \text{ consecutive bits flipped}] = \sum_{i=1}^{n-\log_2 n-k} P[E_i] = \sum_{i=1}^{n-\log_2 n-k} \prod_{t=i}^{i+\log_2 n+k} P[X_i = 1]
\]

\[
(n - \log_2 n - k) \left(\frac{1}{2}\right)^{\log_2 n+k} < \left(\frac{1}{2}\right)^k
\]

2nd inequality follows from Fact 1 about the probability of event union, the 4th equality follows from independence of bit flips.
An often useful theorem

**Theorem**

Assume $E_1, \ldots E_n$ are mutually disjoint events such that $\bigcup_{i=1}^{n} E_i = \emptyset$. Then, considered any event $B$:

$$
P[B] = \sum_{i=1}^{n} P[B \cap E_i] = \sum_{i=1}^{n} P[B \mid E_i] P[E_i]
$$

**Law of total probability**

You should convince yourself (and prove) that the theorem works.

**What happens if the $E_i$’s are not disjoint?**
Discrete random variables

**Definition (Random variable)**

Random variable on a sample space $O$:

$$X : O \rightarrow \mathbb{R}$$

$X$ is **discrete** if it can only take on a finite or countably infinite set of values

**Independence**

$X, Y$ independent if and only if

$$P[(X = x) \cap (Y = y)] = P[X = x]P[Y = y]$$

for all possible values $x, y$

$X_1, \ldots, X_k$ mutually independent if and only if, for every $I \subseteq \{1, \ldots, k\}$ and values $x_i, i \in I$:

$$P[\bigcap_{i \in I}(X_i = x_i)] = \prod_{i \in I} P[X_i = x_i]$$
Expectation of discrete random variables

**Definition (Expectation)**

Random variable $X$ on a sample space $O$.

$$E[X] = \sum_i i P[X = i],$$

where $i$ varies over all possible values in the range of $X$.

**Theorem (Linearity of expectation)**

For any finite collection $X_1, \ldots, X_k$ of discrete random variables:

$$E \left[ \sum_{i=1}^k X_i \right] = \sum_{i=1}^k E[X_i]$$

**Note:** This result holds always.
**Example**

**A:** Assume we toss a fair coin \( n \) times. Let \( X \) denote the number of heads. Determine \( E[X] \).
**Example**

**A**: Assume we toss a fair coin $n$ times. Let $X$ denote the number of heads. Determine $E[X]$.

We define binary variables $X_1, \ldots, X_n$, with $X_i = 1$ if the $i$-th coin toss gave head, 0 otherwise. We obviously have:

$$X = \sum_{i=1}^{n} X_i$$

Hence:

$$E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} P[X_i = 1] = \frac{n}{2}$$
Bernoulli variable and binomial distribution

Assume an experiment succeeds with probability $p$ and fails with probability $1 - p$. The following is Bernoulli indicator variable:

$$Y = \begin{cases} 1, & \text{The experiment succeeds} \\ 0, & \text{Otherwise} \end{cases}$$

Of course: $E[Y] = P[Y = 1] = p$ (prove)

Binomial distribution

Consider $n$ independent trials of the experiment and let $X$ denote the number of successes. Then $X$ follows the binomial distribution:

$$P[X = i] = \binom{n}{i} p^i (1 - p)^{n-i}$$

Q4: prove the claim above. Prove that $E[X] = np$
Geometric distribution

Consider the number $Z$ of independent trials until the first success of the experiment. *Prove* that $X$ follows a geometric distribution with parameter $p$, i.e.:

$$
P[Z = i] = (1 - p)^{i-1} p.
$$

**Expectation**

**Q5**: prove that $E[Z] = \frac{1}{p}$. **Hint.** Use the following result:

**Lemma**

*Assume $Z$ is a discrete random variable that takes on only non-negative values:*

$$
E[Z] = \sum_{i=1}^{\infty} P[Z \geq i]
$$
**Definition**

Assume $X$ and $Y$ are discrete random variables.

\[ E[X \mid Y = i] = \sum_j jP[X = j \mid Y = i], \]

where $j$ varies in the range of $X$.

The following holds:

**Lemma**

\[ E[X] = \sum_i E[X \mid Y = i] P[Y = i], \]

where $i$ varies over the range of $Y$. 
Variance and more...

**Definition**

If $X$ is a random variable

$$\text{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

$\sigma(X) = \sqrt{\text{var}[X]}$ is the *standard deviation* of $X$.

The following holds:

**Lemma**

*If $X_1, \ldots, X_k$ are mutually independent random variables:*

$$\mathbb{E}\left[\prod_{i=1}^{k} X_i\right] = \prod_{i=1}^{k} \mathbb{E}[X_i]$$

Q6: prove the lemma for $k = 2$. 
Example

Assume we observe a binary string $S$ of variable length. In particular, the length of the string falls in the range $\{1, \ldots, n\}$ with uniform probability, while for any particular string length, every bit is 1 or 0 with equal probability, independently of the others. What is the average number of 1’s observed?
Example

Assume we observe a binary string $S$ of variable length. In particular, the length of the string falls in the range $\{1, \ldots, n\}$ with uniform probability, while for any particular string length, every bit is 1 or 0 with equal probability, independently of the others. What is the average number of 1’s observed?

**Sol.:** we apply Lemma 8. More in detail, let $L$ denote the random variable that gives the length of the string. For any fixed value $k$ of $L$, we define binary variables $X_1, \ldots, X_k$, where $X_i$ is equal to the $i$-th bit of the string. If $Y$ denotes the number of 1’s in $S$ have:

$$E[Y \mid L = k] = \sum_{i=1}^{k} P[X_i = 1 \mid L = k] = \frac{k}{2}.$$  

Applying Lemma 8:

$$E[Y] = \sum_{L=1}^{n} E[Y \mid L = k] P[L = k] = \frac{1}{n} \sum_{k=1}^{n} \frac{k}{2} = \frac{n + 1}{4}.$$
“Concentration of measure refers to the phenomenon that a function of a large number of random variables tends to concentrate its values in a relatively narrow range (under certain conditions of smoothness of the function and under certain conditions on the dependence amongst the set of random variables)” [Dubhashi and Panconesi, 2009].

In this lecture

- General but weaker results (Markov’s and Chebyshev’s inequality)
- Strong results for the sum of independent random variables in [0, 1] (Chernoff bound)
Markov’s and Chebyshev’s inequalities

Theorem (Markov’s inequality)

Let $X$ denote a random variable that assumes only non-negative values. Then, for every $a > 0$:

$$
P[X \geq a] \leq \frac{E[X]}{a}.
$$

Theorem (Chebyshev’s inequality)

Let $X$ denote a random variable. Then, for every $a > 0$:

$$
P[|X - E[X]| \geq a] \leq \frac{v_ar[X]}{a^2}.
$$
Markov vs Chebyshev

- Markov inequality applies to *non-negative* variables, while Chebyshev’s to any variable
- Chebyshev’s inequality often stronger, but you need at least upper bound on variance (not always trivial to estimate)

**Example (Markov)**

Consider *n independent* flips of a fair coin. Use Markov’s and Chebyshev’s inequalities to give bound on the probability of obtaining more than $3n/4$ heads.
Markov vs Chebyshev

- Markov inequality applies to non-negative variables, while Chebyshev’s to any variable
- Chebyshev’s inequality often stronger, but you need at least upper bound on variance (not always trivial to estimate)

Example (Markov)

Consider \( n \) independent flips of a fair coin. Use Markov’s and Chebyshev’s inequalities to give bound on the probability of obtaining more than \( 3n/4 \) heads.

**Sol.:** Let \( X_i = 1 \) if \( i \)-th coin toss gives heads 0 otherwise and let \( X = \sum_{i=1}^{n} X_i \). Of course, \( E[x] = n/2 \). Applying Markov’s inequality thus gives:

\[
P \left[ X > \frac{3}{4} n \right] \leq \frac{n/2}{3n/4} = \frac{2}{3}.
\]
Example (Chebyshev)

We need the variance of $X$ in order to apply Chebyshev's inequality. We have:

$$\text{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}\left[\left(\sum_{i=1}^{n}(X_i - \frac{1}{2})^2\right)\right] + 2\sum_{i=1}^{n-1}\sum_{h=i+1}^{n}\mathbb{E}[(X_i - \frac{1}{2})(X_h - \frac{1}{2})]$$

$$= n\sum_{i=1}^{n}\text{var}[X_i] = \frac{n}{4},$$

where last equality follows since i) the $X_i$ are mutually independent, ii) $\mathbb{E}[X_i] = 1/2$ for every $i$ and iii) $\text{var}[X_i] = 1/4$ for every $i$. 

Example (Chebyshev)

We need the variance of $X$ in order to apply Chebyshev’s inequality. We have:

\[
\text{var} [X] = \mathbb{E} [(X - \mathbb{E}[X])^2] = \mathbb{E} \left[ \left( \sum_{i=1}^{n} (X_i - \frac{1}{2}) \right)^2 \right] \\
= \sum_{i=1}^{n} \mathbb{E} \left[ (X_i - \frac{1}{2})^2 \right] + 2 \sum_{i=1}^{n-1} \sum_{h=i+1}^{n} \mathbb{E} \left[ (X_i - \frac{1}{2})(X_h - \frac{1}{2}) \right] \\
= \sum_{i=1}^{n} \text{var} [X_i] = \frac{n}{4},
\]

where last equality follows since i) the $X_i$ are mutually independent, ii) $\mathbb{E}[X_i] = 1/2$ for every $i$ and iii) $\text{var} [X_i] = 1/4$ for every $i$. 
Example (Chebyshev cont.)

Now, from Chebyshev's inequality:
\[ P \left[ X \geq \frac{3}{4} \right] \leq P \left[ |X - E[X]| \geq n \right] \leq \text{var} [X] \left( \frac{n}{4} \right)^2 = 4n. \]

Observe the following:

This result is much stronger than previous one.

We implicitly proved a special case of a general result:

**Theorem (Variance of the sum of independent variables)**

If \( X_1, \ldots, X_n \) are mutually independent random variables:

\[ \text{var} \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} \text{var} [X_i]. \]
Example (Chebyshev cont.)

Now, from Chebyshev’s inequality:

\[ P \left[ X \geq \frac{3}{4} n \right] \leq P \left[ |X - E[X]| \geq \frac{n}{4} \right] \leq \frac{\text{var}[X]}{(n/4)^2} = \frac{4}{n}. \]

Observe the following:

- This result is much stronger than previous one
- We implicitly proved a special case of a general result:
Example (Chebyshev cont.)

Now, from Chebyshev’s inequality:

\[ P \left[ X \geq \frac{3}{4} n \right] \leq P \left[ |X - E[X]| \geq \frac{n}{4} \right] \leq \frac{\text{var} [X]}{(n/4)^2} = \frac{4}{n}. \]

Observe the following:

- This result is much stronger than previous one
- We implicitly proved a special case of a general result:

Theorem (Variance of the sum of independent variables)

If \( X_1, \ldots, X_n \) are mutually independent random variables:

\[ \text{var} \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} \text{var} [X_i]. \]
Poisson trials

Definition

\(X_1, \ldots, X_n\) form a sequence of Poisson trials if they are binary and mutually independent, so that \(P[X_i = 1] = p_i\), \(0 < p_i \leq 1\).

Note the difference with Bernoulli trials: these are the special case of Poisson trials when \(p_i = p\), for every \(i\). In the next slides:

- We assume a sequence \(X_1, \ldots, X_n\) of independent Poisson trials
- In particular: \(P[X_i = 1] = p_i\)
- \(X = \sum_{i=1}^{n} X_i\) and \(\mu = E[X]\).
Chernoff bound(s)

A set of powerful concentration bounds. Hold for the sum or linear combination of Poisson trials.

Theorem (Chernoff bound (upper tail)[Mitzenmacher and Upfal, 2005])

Assume $X_1, \ldots, X_n$ form a sequence of independent Poisson trials, so that $P[X_i = 1] = p_i$, $X = \sum_{i=1}^{n} X_i$ and $\mu = E[X]$. Then:

For $\delta > 0$: $P[X \geq (1 + \delta)\mu] < \left(\frac{e^\delta}{(1 + \delta)^{1+\delta}}\right)^\mu$ (1)

For $0 < \delta \leq 1$: $P[X \geq (1 + \delta)\mu] \leq e^{-\frac{\delta^2}{3}\mu}$ (2)

For any $t \geq 6\mu$: $P[X \geq t] \leq 2^{-t}$ (3)
Theorem (Chernoff bound (lower tail) [Mitzenmacher and Upfal, 2005])

Under the same assumptions, for $0 < \delta < 1$:

$$P[X \leq (1 - \delta)\mu] < \left( \frac{e^{-\delta}}{(1 - \delta)^{1-\delta}} \right)^\mu$$  \hspace{1cm} (4)

For $0 < \delta \leq 1$: $P[X \leq (1 - \delta)\mu] \leq e^{-\frac{\delta^2}{2}\mu}$  \hspace{1cm} (5)

- (2) and (5) most used in practice
- Many different versions of the bound exist for different scenarios, also addressing the issue of (limited) dependence [Mitzenmacher and Upfal, 2005, Dubhashi and Panconesi, 2009]
Sketch of proof for Chernoff bounds

Consider the upper tail. The proof uses Markov’s inequality in a very smart way. In particular, considered any $s > 0$:

$$
P[X \geq (1 + \delta)\mu] = \mathbb{P}[e^{sX} \geq e^{s(1+\delta)\mu}] \leq \frac{\mathbb{E}[e^{sX}]}{e^{s(1+\delta)\mu}}$$

$$= \frac{\prod_{i=1}^{n} \mathbb{E}[e^{sX_i}]}{e^{s(1+\delta)\mu}} = \frac{\prod_{i=1}^{n} (1 + p_i(e^{s} - 1))}{e^{s(1+\delta)\mu}} \leq \frac{\prod_{i=1}^{n} e^{p_i(e^{s} - 1)}}{e^{s(1+\delta)\mu}}$$

$$= \frac{e^{(e^{s} - 1)\mu}}{e^{s(1+\delta)\mu}}.$$

- Second inequality follows from Markov’s inequality, third equality from independence of the $X_i$’s, fourth inequality since $1 + x \leq e^x$
- Bounds follow by appropriately choosing $s$ (i.e., optimizing w.r.t. $s$)
Example: coin flips

\( X \) (no. heads) the sum of independent Poisson trials (Bernoulli trials in this case), with \( \mathbf{E}[X] = n/2 \). We apply bound (2) with \( \delta = 1/2 \) to get:

\[
P \left[ X \geq \frac{3}{4} n \right] = P \left[ X \geq (1 + \delta) \mathbf{E}[X] \right] \leq e^{-\frac{n}{12}}
\]

Remarks

- Useful if \( n \) large enough
- Observe that \( P[X \geq 3n/4] \)
  - \( \leq 2/3 \) (Markov)
  - \( \leq 4/n \) (Chebyshev)
  - \( \leq e^{-\frac{n}{12}} \) (Chernoff)
- Concentration results at the basis of statistics
Dictionaries

A dynamic set $S$ of objects from a discrete universe $U$, on which (at least) the following operations are possible:

- Item insertion
- Item deletion
- Set membership: decide whether item $x \in S$

Typically, it is assumed that each item in $S$ is *uniquely* identified by a *key*. Let $\text{obj}(k)$ be item with key $k$:

**Operations**

- $\text{insert}(x, S)$: insert item $x$
- $\text{delete}(k, S)$: delete $\text{obj}(k)$
- $\text{retrieve}(k, S)$: retrieve $\text{obj}(k)$

This is a minimal set of operations. Any database implements a (greatly augmented) dictionary
**Hash functions**

- Often used to implement insert, delete and retrieve in a dictionary
- In general, a hash function $h : U \rightarrow [n]$ maps elements of some discrete universe $U$ onto integers belonging to some range $[n] = \{0, 1, \ldots, n - 1\}$. Typically, $|U| >> n$. Ideally, the mapping should be uniform. We assume without loss of generality that $U$ is some subset of the integers (why can we state this?)

Ideal behaviour: items in $U$ mapped uniformly at random in $\{0, \ldots, n-1\}$
Hash functions/2

Mapping should look “random” → If $m$ items are mapped, then every $i \in [n]$ should be the image of $\approx m/n$ items.

- E.g.: if $U = \{0, \ldots, m - 1\}$ consider $h(x) = x \mod p$, with $p$ a suitable prime.

- Problem: this works if items from $U$ appear at random → often many correlations present

- **Q7a:** create an adversarial sequence that maps all elements of the sequence onto the same $i$

- **Q7b:** Assume $n \leq m$ items chosen u.a.r. from $U$ are inserted into a hash table of size $p$, using the hash function $h(x) = x \mod p$, with $p$ a suitable prime. What is the expected number of items hashed to the same location of the hash table?
Randomizing the hash function

Use a randomly generated hash function to map items to integers.

**Idea:** even if correlations present, items are mapped randomly. Ideal behaviour

- For each \( x \in U \), \( P[h(x) = j] = 1/n \), for every \( j = 1, \ldots, n \)
- The values \( h(x) \) are *independent*

**Caveats**

- This does not mean that every evaluation of \( h(x) \) yields a different random mapping, but only that \( h(x) \) is equally likely to take any value in \([0, \ldots, n - 1]\)
- Not easy to design an “ideal” hash function (many truly random bits necessary)
Families of universal hash functions

We assume we have a suitably defined family $\mathcal{F}$ of hash functions, such that every member of $h \in \mathcal{F}$ is a function $h : U \rightarrow [n]$.

**Definition**

$\mathcal{F}$ is a 2-universal hash family if, for any $h(\cdot)$ chosen *uniformly at random* from $\mathcal{F}$ and for every $x, y \in U$ we have:

$$P[h(x) = h(y)] \leq \frac{1}{n}.$$

- Definitions generalizes to $k$-universality
  [Mitzenmacher and Upfal, 2005, Section 13.3]
- **Problem**: define “compact” universal hash families
A 2-universal family

Assume $U = [m]$ and assume the range of the hash functions we use is $[n]$, where $m \geq n$ (typically, $m \gg n$). We consider the family $\mathcal{F}$ defined by $h_{ab}(x) = ((ax + b) \mod p) \mod n$, where $a \in \{1, \ldots, p - 1\}$, $b \in \{0, \ldots, p\}$ and $p$ is a prime $p \geq m$.

How to choose u.a.r. from $\mathcal{F}$

For a given $p$: Simply choose $a$ u.a.r. from $\{1, \ldots, p - 1\}$ and $b$ u.a.r. from $\{0, \ldots, p\}$.
A 2-universal family/cont.

Theorem ([Carter and Wegman, 1979, Mitzenmacher and Upfal, 2005])

\( \mathcal{F} \) is a 2-universal hash family. In particular, if \( a, b \) are chosen uniformly at random:

\[
P[h_{ab}(x) = i] = \frac{1}{n}, \forall x \in U, i \in [n].
\]

\[
P[h_{ab}(x) = h_{ab}(y)] \leq \frac{1}{n}, \forall x, y \in U.
\]
Example: hash tables

Consider a hash table implemented as follows:

- An array $A$ of lists of size $n$
- $h : U \rightarrow [n]$, mapping each object in $U$ onto a position of $A$
- $A_i$ is the list of objects hashed to position $i$ (collisions solved by concatenation)
Case 1

Assume that \( h(\cdot) \) is selected uniformly at random from an “ideal” family, so that:

1. \( P[h(x) = i] = \frac{1}{n}, \forall x \in U, i \in [n] \)
2. \( \forall k, x_1, \ldots, x_k \in U, \forall y_1, \ldots, y_k \in [n] \):

\[
P \left[ \bigcap_{i=1}^{k} (h(x_i) = y_i) \right] = \prod_{i=1}^{k} P[h(x_i) = y_i] = \frac{1}{n^k}
\]

Q8

Consider the insertion of the \( m \) elements of \( U \) and denote by \( S_i \) the size of list \( A_i \). Prove the following: for \( 0 < \epsilon < 1 \)

\[
P \left[ \exists i : S_i > (1 + \epsilon) \frac{m}{n} \right] \leq \frac{1}{n},
\]

whenever \( m = \Omega \left( \frac{1}{\epsilon^2} n \ln n \right) \) (Use Chernoff bound)
Case 2
Assume that $h(\cdot)$ is selected uniformly at random from a 2-universal hash family

Q9
Prove that the following, much weaker result holds:

$$
\Pr \left[ \exists i : S_i \geq m \sqrt{\frac{2}{n}} \right] \leq \frac{1}{2}.
$$

Hints:

1. Define $X_{jk} = 1$ iff items $j$ and $k$ mapped onto same array position and let $X = \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} X_{jk}$ the total number of collisions.

2. Note that, if the maximum number of items mapped to the same position in $A$ is $Y$, then $X \geq \binom{Y}{2}$


Internet traffic growth: sources and implications.