COSC 290 Discrete Structures

Lecture 36: Bloom filters & Randomized Response

Prof. Michael Hay
Friday, May 4, 2018
Colgate University

Plan for today

1. Bloom filters
2. Expectation
3. Randomized response

Bloom filters

Bloom filters

Purpose: keep track of which objects you have seen before.

- Insert(x): inserts x into bloom filter
- Lookup(x): returns True if x appears to have been inserted before, False otherwise

Bloom filter:

- A of size m. Each cell stores a bit.
- k hash functions: $h_1, h_2, \ldots, h_k$.
- Each hash function maps object x to cell in A.

Insert(x): compute $h_1(x), h_2(x), \ldots, h_k(x)$. Set those bits to 1 in A.

Lookup(x): compute $h_1(x), h_2(x), \ldots, h_k(x)$. Check these cells in array A. If they are *all* equal to 1, then return True. Otherwise false.

When bloom filters work well

Bloom filters sometimes produce *false positives*.

How do avoid false positives:

- Make array A large. The larger A is, the less likely a hash collision.
- Set k to be "just right":
  - When k is too small, a few collisions can cause a false positive. (Imagine $k = 1$)
  - When k is too big, each insertion flips a lot of bits and we quickly "run out of bits." (Imagine $k = m$)

Strongly encourage you to look at exercises DLN 10.99-10.104. Useful study for the final exam!

Expectation

Ferengi population

Last time we looked at B, the number of boys, and G, the number of girls and we found that:

- $\mathbb{E}[B] = 1$
- $\mathbb{E}[G] = 1$

Let's define a new random variable, $T := B + G$ (total number of children). What is $\mathbb{E}[T]$?

Poll: expected value of T

Recall that $\mathbb{E}[B] = \mathbb{E}[G] = 1$. Let $T := B + G$ (total number of children). What is $\mathbb{E}[T]$?

$$
\begin{aligned}
\mathbb{E}[T] &= \sum_{s \in S} Pr(s) \cdot T(s) && \text{definition of expectation} \\
&= \sum_{s \in S} Pr(s) \cdot (B(s) + G(s)) && \text{definition of } T \\
&= \sum_{s \in S} Pr(s) \cdot B(s) + Pr(s) \cdot G(s) && \text{distribute mult. over addition} \\
&= \sum_{s \in S} Pr(s) \cdot B(s) + \sum_{s \in S} Pr(s) \cdot G(s) && \text{split into two summations} \\
&= ??? && \text{Can you simplify last line and solve?}
\end{aligned}
$$

Therefore, $\mathbb{E}[T]$ is (a) 0, (b) 1, (c) 1.5, (d) 2, (e) 2.5.

Linearity of expectations

Let X_1 and X_2 be any two random variables.

$$\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$$

More generally, let $X_1, X_2, \ldots, X_n$ be any n random variables.

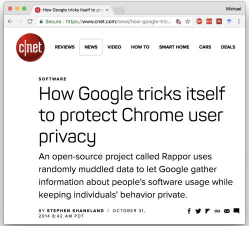
$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

Finally, it's not hard to show that for any a that is *constant*,

$$\mathbb{E}[aX] = a\mathbb{E}[X]$$

Randomized response

Using randomization to safely extract private information


How Google tricks itself to pro

Michael

Secure https://www.cnet.com/news/how-google-tric...

c|net REVIEWS NEWS VIDEO HOW TO SMART HOME CARS DEALS

SOFTWARE

How Google tricks itself to protect Chrome user privacy

An open-source project called Rappor uses randomly muddled data to let Google gather information about people's software usage while keeping individuals' behavior private.

BY STEPHEN SHANKLAND / OCTOBER 31, 2014 8:42 AM PDT

Privacy through randomization

Suppose pollster wants to ask sensitive question.

Example: Do you support legalization of marijuana? Respondent may be reluctant to answer "Yes."

Randomized response (Warner, 1965)

- Pollster has a *biased* coin: heads with probability p where $p > \frac{1}{2}$. Pollster knows the value of p.
- Pollster asks question. *True response* is kept secret.
- Respondent flips coin. Result of coin flip hidden from pollster.
- Randomized response: If heads, answers truthfully; if tails, lies.
- Respondent tells pollster only the *randomized response*.

(Quick demonstration.)

Poll: probability of a random Yes

Let θ be fraction of population whose *true response* is Yes. Let p be probability of heads.

Suppose a respondent is randomly chosen from population. Let E be the event that the randomly selected respondent gives a *randomized response* of Yes.

What is $Pr(E)$?

A) θ
B) p
C) $\theta \cdot p$
D) $\theta \cdot p + (1-\theta) \cdot (1-p)$
E) Unknown because we don't know *true response*

Indicator random variable

An indicator random variable is a binary random variable (i.e. it maps each outcome to either 0 or 1).

Assume pollster asks n respondents. Each respondent randomly selected from population.

Let X_i be the following indicator random variable,

$$X_i = \begin{cases} 1 & \text{if randomized response of } i^{th} \text{ respondent is "Yes"} \\ 0 & \text{if randomized response of } i^{th} \text{ respondent is "No"} \end{cases}$$

What is $Pr(X_i = 1)$?

$$Pr(X_i = 1) = \theta \cdot p + (1-\theta) \cdot (1-p)$$

What is $Pr(X_i = 0)$?

$$Pr(X_i = 1) = \theta \cdot (1-p) + (1-\theta) \cdot p$$

What can we learn about θ?

Suppose we repeat this process with a sample of n respondents. Let $Y := \sum_{i=1}^{n} X_i$.

What is $\mathbb{E}[Y]$?

- Use linearity of expectations: $\mathbb{E}[Y] = \mathbb{E}[\sum_i X_i] = \sum_i \mathbb{E}[X_i]$.
- $\mathbb{E}[X_i] = Pr(X_i = 1)$

Let's rearrange and "solve" for θ:

$$\theta = \frac{\frac{\mathbb{E}[Y]}{n} - (1-p)}{(2p-1)}$$

Key point: if you knew $\mathbb{E}[Y]$, you'd know θ. Unfortunately we don't know $\mathbb{E}[Y]$. However, we can *estimate* it!

Estimating θ

Let $\hat{\theta}$ denote the following random variable

$$\hat{\theta} := \frac{\frac{Y}{n} - (1-p)}{(2p-1)}$$

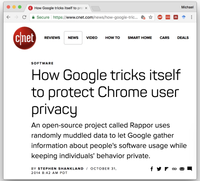
What is $\mathbb{E}\left[\hat{\theta}\right]$? $\mathbb{E}\left[\hat{\theta}\right] = \theta$ (an unbiased estimator)

How accurate is $\hat{\theta}$? We can look at the *variance* of $\hat{\theta}$, which is a measure of how much it deviates from its expected value.

$$V(\hat{\theta}) = \underbrace{\frac{\theta(1-\theta)}{n}}_{\text{error from sampling}} + \underbrace{\frac{p(1-p)}{n(2p-1)^2}}_{\text{error due to randomized answers}}$$

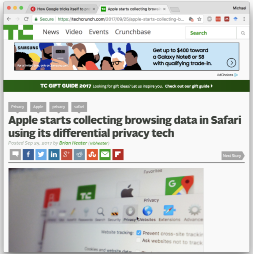
What happens when $p = 1/2$? $p = 1$? $p = 0$? (Note: You can derive this result using definition of V in book and the fact that $V(\sum_i X_i) = \sum_i V(X_i)$ when X_i are independent, which they are here.)

Using randomization to safely extract private information



Google's approach: compress user data using bloom filter, then use randomized response on each bit of bloom filter.

Apple uses similar technologies



Exercise

Consider this alternative randomized protocol.

Flip coin: if heads, answer Yes; if tails, answer truthfully.

What is $\mathbb{E}[Y]$ under this randomized model?

As before,

- assume θ fraction of the population would answer Yes truthfully.
- use linearity of expectations: $\mathbb{E}[\sum_i X_i] = \sum_i \mathbb{E}[X_i]$
- for indicator random variable $\mathbb{E}[X_i] = Pr(X_i = 1)$

Does this approach leak more/less information than previous approach?