Topics in Data Science, Data Protection — Spring 2020
CS 4364 — L. Longpré

Class notes 2/25/2020

1 Huffman codes

Exercise 4 from last class:

The following is a representation of a code using the algorithm seen in class. The possible characters are a, b, c, d, e, f, g, each represented by 3 bits. What is the code?
00100010110010010110111100101011 00

Exercise

Suppose we build a Huffman code as follows. Compute the frequency of each possible byte from a file. Build a Huffman code for each of the byte with a non-zero frequency. Suppose there are n such bytes. Compute a representation of the code using the method seen in class, but putting the actual byte on leaves of the Huffman code tree. What is the size of the representation in bytes?

2 Entropy Review

(Refer to Denning's book, section 1.4.1.)
Let $X_1, ..., X_n$ be n possible messages occurring with probabilities $p(X_1), ..., p(X_n)$, where $\sum_{i=1}^{n} p(X_i) = 1$. The *entropy* of a given message is defined by the weighted average:

$$H(X) = -\sum_{i=1}^{n} p(X_i) \lg p(X_i) = \sum_{i=1}^{n} p(X_i) \lg \left(\frac{1}{p(X_i)} \right).$$

Conditional entropy of X given Y:

$$H_Y(X) = \sum_{X,Y} p_Y(X) p(Y) \lg \left(\frac{1}{p_Y(X)} \right). \quad (1)$$

$$= \sum_{Y} p(Y) \sum_{X} p_Y(X) \lg \left(\frac{1}{p_Y(X)} \right). \quad (2)$$

3 Encryption and perfect secrecy

Consider an encryption system with possible encryption keys K, possible messages M and possible encrypted messages C. The encryption system has perfect secrecy if $p_C(M) = p(M)$.
Examples of encryption systems: Cæsar, one-time pad, RSA. Which one(s) has perfect secrecy?

3.1 Perfect secrecy in statistical database

Definition?
In practice, no statistical database can provide perfect secrecy.

3.2 Macrostatistics

Tables with collection of related statistics, like counts and sums.

3.3 Microstatistics

Statistical evaluation programs used to compute statistics. Protection mechanisms applied before statistics are published.

3.3.1 Census bureaus disclosure control

- Remove identifying information from records
- Add noise to data
- suppress sensitive data
- remove records with extreme values
- statistics based an relatively small samples of complete data

3.4 Query Processing Systems

(We covered this in class already.)

4 Attacks

4.1 Trackers

(We covered this in class already.)

4.2 Linear System Attacks

Example with key specified queries.

Approach still possible for characteristic specified queries.

4.3 Median attacks

4.4 Insertion and deletion attacks

5 Control mechanisms

5.1 Maximum-order control

Restrict queries that employs too many attribute values.

5.2 Cell Suppression

Suppress cells containing sensitive data, like data based on too few individuals, like an n-respondent, $k\%$ dominance rule. We may need to suppress some nonsensitive statistical cells, called *complementary suppressions*.

Example

Counts

Sex	1978	1979	1980	1981	Sum
Female	1	2	2	1	6
Male	3	2	0	2	7
Sum	4	4	2	3	13

Total SAT scores

Sex	1978	1979	1980	1981	Sum
Female	800	1330	1120	500	3750
Male	1930	1150	0	1180	4260
Sum	2730	2480	1120	1680	8010

Total SAT scores

Sex	1978	1979	1980	1981	Sum
Female	—	1330	1120	—	3750
Male	1930	1150	0	1180	4260
Sum	2730	2480	1120	1680	8010

Table with non-negative values

x_{11}	6	x_{13}	25
8	x_{22}	x_{23}	30
x_{31}	x_{32}	3	20
20	30	25	75

0..12	6	7..19	25
8	7..19	3..15	30
0..12	5..17	3	20
20	30	25	75