Privacy pangenomics

Christos Chatzifountas

December 11, 2021

Introduction

Haplotype database representation

Objective

We want to apply differential privacy Techniques to haplotypes

Basics

A database x is a collection of elements of a universe $\mathcal{X}$ of rows (records) The histogram of a database is a vector $x_i \dots x_n, n = |\mathcal{X}|$and x_i is the number of repeats of a row in the database. The l_1 norm of a database is defined in thought of it's histogram:

$$\|x\|_1 = \sum_{i=1}^{|\mathcal{X}|} |x_i| \tag{1}$$

The distance is defined thought the norm in the usual way

A haplotype is a collection of nodes and edges (connections between nodes), in other words a graph database. Graph databases can be represented relational through with adjacency lists So for any graph we need a column of nodes, and a column of adjacent nodes

Lemma 1. *Let x,y be haplotypes. If they differ in a node then* $\|x - y\|_1 \geq 1$

Proof. Let x,y be haplotypes that on one node or more. Then that node and the adjacency differ in their adjacency list representation. Hence their

histograms differ on two or more coordinates. If two integers are different, their difference is greater than one , hence $\|x-y\| \geq |x_n - y_n| + |x_m - y_m| \geq 2$ □

Db selection

In a graph database the l_1 norm is not very natural and is the usual norm used in the context of differential privacy, so is sensible for the current applications to switch to a row database and have each haplotype to be a "record" in our universe. For a database x of in that universe the Range of the utility function is all the pairs (h, n), where h is a haplotype and n is a natural number. Then $u(x, (h, n))$ should output maximum utility when $n = \phi_x(h)$, n is equal to the number of times that h appears in x. Also we have to take in to account the dimension of the histogram of x. Recall that a histogram is a vector of the form $(x_1, x_2, x_3 \ldots)$ where x_n is the number of appearances of a haplotype in x.

Definition 0.1 (Utility function)**.** Let x be a collection of haplotypes, $\phi(h)$ the frequency of $h \in x$ and $l(h)$ the length of h. Then we define the utility of a haplotype-frequency pair $r = (h, n)$, $n \in \mathbb{N}$ as

$$u(x, r) = \frac{\ln(\phi(h) + 1)}{1 + (\phi(h) - n)^2} \sqrt{\frac{l(h)}{\alpha + l(h)}} \tag{2}$$

Where α is a (suitable) constant