CS 6993 Independent Study 8 May 2018

Privacy and Security in Bioinformatics

Author: Nishant Jha *Advisor: Mohammad Mahmoody*

1 Garbled Circuits

Motivation

Setup

Three Securiy Notions

1. **Obliviousness**
2. **Privacy**
3. **Authenticity**

Generation

Evaluation

Demo

2 Private Edit Distance

Introduction The Human Genome contains around 3,200,000,000 base pairs, often denoted as 3,200 Mb. [1] Each base pair can have one of four values: (A, T, C, G) and so two bits are sufficient to represent all base pairs.

```
A => 00
T => 01
C => 10
G => 11
```

This means a single byte can represent a sequence of 4 base-pairs. Consequently, the Human Genome is about 800 megabytes long. Unfortunately, this is merely a theoretical lower bound. In practice, the average sequenced genome occupies around 200 gigabytes of disk space. This is a byproduct of how modern genetic sequencers work and the file format used to store the resulting data.

This unwieldy size necessitated an optimization that takes advantage of the large percentage of the genome sequence that is conserved between individual humans. If a comparison needs to be made between two different genome sequences, a scientist can simply analyze how each genome sequence differs from a *reference genome sequence* rather than directly compare the base-pair sequences for each. This allows for a significant reduction in size. A fresh-off-the-sequencer genome is usually stored in a 200 GB FASTQ file, while the "diff" optimized storage method is stored in a variant (".vcf") file. This ".vcf" file usually only occupies around 125 MB which is more space efficient by orders of magnitude.

An analogous idea in computer science is that of *edit distance*. Put simply, the edit distance between two strings is the number of characters that must be substituted, inserted, or deleted to change one of the strings into another.

It is often useful to calculate the edit distance between two different genomes. If a patient is diagnosed with a disease, a physician may want to see the prognosis of other genetically similar patients with the disease.

From both a practical and privacy standpoint, it is undesirable to send a patient's entire genome sequence whenever a comparison is required. Because each genome sequence is 200 gigabytes in size, bandwidth will quickly become a bottleneck. Additionally, genome sequences are practically unparalleled in their ability to uniquely identify an individual. Furthermore, they contain sensitive medical data - not only the prescence of congenital diseases but predispositions to a wide swath of additional maladies.

The .vcf "diff" method only alleviates one of these problems - the one concerning bandwidth. Because the human reference genome sequence is publicly available, reverse engineering an individual's genome sequence given a comprehensive variant file is trivial - simply start with the reference and make the changes described in the variant file. Similarily, if you are given a variant file that describes the set difference between your genome sequence and another person Y's sequence, the derivation of Y's sequence is trivial as well.

Problem Statement Clearly, we cannot allow this set difference to be calculated. However, if we take advantage of three features we can reduce the problem of calculating the set difference while maintaining privacy to the problem of calculating the *size* of the set difference while maintaining privacy.

1. Most differences between human genome sequences are *substitutions*, not insertions or deletions.

2. There exists a public reference genome and many variations that have already been computed.

3. Using probabilistic algorithms, the size of the set difference can be securely approximated while keeping the actual set difference private.

The algorithm is as follows:

1. Party A calculates the minimum edit sequences from the reference genome to genome A using the levenshtein distance.
2. Party B calculates the minimum edit sequences from the reference genome to genome B using the levenshtein distance.
3. The parties run a secure computation protocol to approximate the cardinality of the set difference of the minimum edit sequences.

This third step has it's own multistep protocol. While the technical details can be found in the implementation included with this report, at a high level the protocol works by "squeezing" each set of edit sequences into an integer. This "squeezing" is performed by taking the sum of the binary hash ($h : edit \to \{-1, 1\}$) of each edit in the set. Once the "squeezed" values d_A and d_B are computed, square of the difference is computed $(d_A - d_B)^2$. This squeezing process is performed l times, and then the k-median is taken. Here, l and k are used to bound the accuracy and the efficiency.

The above algorithm is implemented in `sec_ped.py` contained in the `demos` directory (where all other demos will be located). As a comparison for the accuracy of the secure computation demo, an unsecure algorithm is implemented in `unsec_ped.py`.

This approximation process's accuracy is aided by the special distribution of human genome sequences.

For any two suitably unrelated individuals:

1. An overwhelming majority of their sequences are conserved, at least 99%
2. The places at which they differ, i.e. the location of their edits from a reference genome are not close together.
3. A majority of the differences between genome sequences are substitutions, around 80%

3 Differential Privacy

Introduction In the previous section, the utility and risks of sharing an individual's genome, was discussed. This discussion applies to personal medical information in general, whether it be the gut microbiome, the proteome, or simply a medical history. When collecting, storing, and querying this personal data, the privacy of the individuals must be maintained, especially as the ubiquity of collection and analysis increases. More specifically, it is a vital requirement that systems that store sensitive data do not allow adversaries to learn about the prescence or even abscence of an individual's data in a given dataset.

Some Straw-man Solutions To convey the non-triviality of this problem, two common-sense approaches will be briefly introduced and examined.

1. **Only Group Queries** If individual privacy is the concern, why not only allow group-level queries? As an example, consider the group-level query: "How many patients in this ward have diabetes?" An easy way to violate the privacy of patient A is to follow up the previous query with: "How many patients aside from patient A have diabetes?" If the number from the first query is greater than the number from the second, patient A must have diabetes, and must not otherwise. Clearly, the privacy of patient A has been compromised.

2. **Random Noise** Why not obfuscate the data returned by introducing random noise to the output? With medical data, there is always the risk that a random perturbation would lead to an incorrect diagnosis or treatment. Additionally, the true value of the output could be determined statistically using repeated queries.

Definition In order to propose a solution that is superior to those described above, the problem must be formally stated. At a high level, Differential Privacy is the requirement that the result of a query should not reveal the prescence or abscence of any individual record in the input dataset. Informally, the outcome of a query should be nearly equally likely if the dataset has your information in it, as to if it did not.

Formal Definition Formally, given a dataset A and B, A and B are said to be *adjacent* if there is a single record r that is in one but not the other. For adjacent datasets A and B, a query Q has ϵ-differential privacy if for any adjacent datasets A and B, and any subset C of possible outcomes $Range(Q)$, $\Pr[Q(A) \in C] \leq \exp(\epsilon) \times \Pr[M(B) \in C]$

Laplace Noise A function f can be made to satisfy ϵ-differential privacy by adding Laplace noise to it. This noise takes the form of a variable L that follows the Laplace Distribution.

$f(x) + L$ where $L = Laplace(0, \sigma)$, $\sigma \geq \frac{\Delta f}{\epsilon}$, and $\Delta f = \max(||f(x) - f(x')||_1)$.

The PDF for the Laplace Distribution is defined as: $\frac{1}{2b} \times \exp(-\frac{|x-\mu|}{b})$ where b is a scaling factor. The Manhattan norm is defined as: $||x|_1 = \frac{1}{n}\sum_1^n |x_1|$.

As an example, consider the function f that computes the average weight of a patient cohort. Now consider two adjacent datasets of size 4 $P1$ and $P2$ that differ by only a single patient's data. $P1 = \{110, 130, 145, 210\}$ and $P2 = \{110, 130, 145, 189\}$ We will calculate the average weight of $P1$ while ensuring ϵ-differential privacy. As the minimum weight is 110 and the maximum is 210, $\Delta f = \frac{210-110}{4} = 25$. Adding the term L, where L is the result of sampling $Laplace(0, \frac{25}{\epsilon})$, would preserve the privacy of the patient that is in $P1$ but not $P2$.

Extending Differential Privacy Any algorithm that satisfies ϵ-differential privacy can be augmented to ensure privacy for a group of size s simply by scaling the privacy budget ϵ by s.

Privacy in Pharmacogenetics

References

[1] https://www.nature.com/articles/35057062

[2] https://www.genome.gov/19016904/faq-about-genetic-and-genomic-science/

[3] http://www.internationalgenome.org/wiki/Analysis/vcf4.0

[4] http://repositorio.uchile.cl/bitstream/handle/2250/126168/Navarro_Gonzalo-_Guided_tour.pdf