Mitigating privacy risks in sharing summary statistics of clinical notes from distributed sources
Software Manual

1 System and Software Requirements

This software has been tested using Google cloud's `n1-standard-8` (Ubuntu 16.04.3 LTS) and Amazon's `r3.xlarge` (Ubuntu 16.04.2 LTS). However, this package should work in lower configuration machines as long as these machines have multiple cores.
The following software modules are required. An installation script (`script.sh`) is provided with the package to install these modules.

- libssl
- NFLlib
- GMP
- MPFR
- OpenMP

GCC 5.4.1 or later is recommended.

2 Entities in the System Architecture and Corresponding Executable

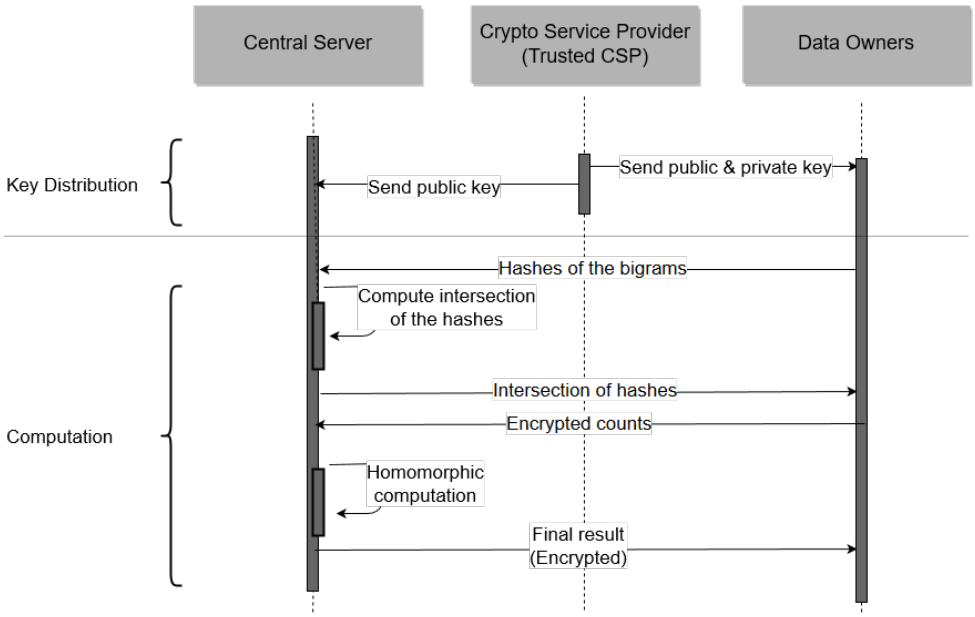

Figure 1: Interaction among different entities of the system.

Table 1: Entity and corresponding executable

Entity	Executable
Crypto Service Provider (CSP)	csp.cpp
Central Server	central_server.cpp
Data Owner	data_owner.cpp

3 Directory Structure

Common directory and files:

- `common`: directory containing some common utilities
- `include`: directory containing nfl source
- `lib`: directory containing nfl source
- `FV_2.hpp`: library for FV cryptosystem
- `fvnamespace2.h`: for FV cryptosystem
- `script.sh`: script for dependency installation

Entity specific directory and files are described in the following subsections.

3.1 Central Server

- `central_server`: directory for central server files
- `central_server.cpp`

3.2 CSP

- `csp`: directory for CSP files
- `csp.cpp`

3.3 Data Owner

- `data_owner_id`, (id = 1, 2, 3...): directory for data owner files. This directory contains some sub-directories for organizing datasets used in different executions of the system protocol. Each sub-directory is identified by an experiment code (A, B, C etc.). This experiment code can be specified in the configuration file (please see Section 5).
- `data_owner.cpp`

4 Compilation

Compiling the executable for central server:

```
g++ central_server.cpp -o cs.o a.out -std=c++11 -fopenmp  -I./include/ -I./include/nfl/
-I./include/nfl/prng/ -I./lib/prng/ -I./lib/params/ -I./include/nfl/opt/arch/ -lgmpxx -lgmp
 -lmpfr -m64 -DNTT_AVX -DNTT_SSE
```

Compiling the executable for CSP:

```
g++ csp.cpp -o csp.o a.out -std=c++11  -I./include/ -I./include/nfl/ -I./include/nfl/prng/
-I./lib/prng/ -I./lib/params/ -I./include/nfl/opt/arch/ -lgmpxx -lgmp  -lmpfr -m64 -DNTT_AVX
-DNTT_SSE
```

Compiling the executable for data owner:

```
g++ data_owner.cpp -o do.o a.out -std=c++11 -lcrypto  -I./include/ -I./include/nfl/ -I./include/nfl
-I./lib/prng/ -I./lib/params/ -I./include/nfl/opt/arch/ -lgmpxx -lgmp  -lmpfr -m64 -DNTT_AVX
-DNTT_SSE
```

5 Configuration File

At the beginning of the system protocol, as part of system initialization process, the central server sends a configuration file to each party. This configuration file contains IP address and port of the parties. This configuration file should be placed in the central server's directory (`central\_server/config.ini`). A sample configuration file is provided with this package. If software is deployed in multiple machines, then IP addresses should be changed accordingly (Please see Section 9 for details).

6 Format of Input Data

This software expects the input data of data owner in the following format:

```
bigram_1 = count_1
bigram_2 = count_2
bigram_3 = count_3
...................
```

For instance, blood group = 21.
The filename containing the input data should be `ExperimentCode_data owner id_pt_count_map.txt` For instance, for experiment code A, data owner id 1, filename will be `A_1_pt_count_map.txt`. Also, this file will be placed in data owner 1's directory for experiment code A (`data\_owner1/A/`).

7 Output for Participating Data Owners

The final output is stored in the `plaintext_output_bigrams.txt` file in the data owners' directory. This file contains the bigrams meeting the threshold criteria.

8 Running the Software

The central server executable should run at the end because it will expect all other entities to be ready for receiving configuration file. For example, for three data owners, the following sequence can be followed (in five different terminals).

```
./do.o 1
./do.o 2
./do.o 3
./csp.o
./cs.o
```

It should be noted that ID of a data owner is passed as a command line argument.

9 Deploying the Software in Multiple Machines

To deploy each entity in different machine, the executable and the directory should be copied to that machine. For instance, to deploy central server in a different machine, only common files and `central_server.cpp`, `central_server` directory are required. Similarly, for data owner 1, only common files, `data_owner.cpp`, and `data_owner` directory are required.

10 Supplementary Codes

The default number of data owners is three. `src_additional` directory contains codes to handle number of data owners other than three.

Access to Source Code and Contact Information

Source code is available at github.com/Nazmus-Sadat/th_mpsi.
For any information regarding this software, email can be sent to: nsadat@ucsd.edu

License

GPLv3

This software uses the following projects (licensed under GPLv3):

1. NFLlib
2. FV-NFLlib