Mestrado em Segurança Informática
Segurança e Privacidade 2022/2023

Practical Assignment #2

Secure Multiparty Computation

Inês Alexandra Barbosa Fonseca (2019223445)
inesfonseca@student.dei.uc.pt

Conteúdo

1 Introdução **2**

2 Problema **2**

3 Datasets **2**

4 Protocolos **3**
4.1 The naive hashing protocol . . . 3
4.1.1 dataset_M1_5000 & dataset_M2_5000 . . . 3
4.1.2 dataset_M1_10000 & dataset_M2_10000 . . . 3
4.1.3 dataset_M1_50000 & dataset_M2_50000 . . . 4
4.1.4 dataset_M1_100000 & dataset_M2_100000 . . . 4
4.1.5 dataset_M1_150000 & dataset_M2_150000 . . . 5
4.2 The server-aided protocol . . . 6
4.2.1 dataset_M1_5000 & dataset_M2_5000 . . . 6
4.2.2 dataset_M1_10000 & dataset_M2_10000 . . . 6
4.2.3 dataset_M1_50000 & dataset_M2_50000 . . . 7
4.2.4 dataset_M1_100000 & dataset_M2_100000 . . . 7
4.2.5 dataset_M1_150000 & dataset_M2_150000 . . . 8
4.3 The Diffie-Hellman-based PSI protocol . . . 9
4.3.1 dataset_M1_5000 & dataset_M2_5000 . . . 9
4.3.2 dataset_M1_10000 & dataset_M2_10000 . . . 9
4.3.3 dataset_M1_50000 & dataset_M2_50000 . . . 10
4.3.4 dataset_M1_100000 & dataset_M2_100000 . . . 10
4.3.5 dataset_M1_150000 & dataset_M2_150000 . . . 11
4.4 The OT-based PSI protocol . . . 12
4.4.1 dataset_M1_5000 & dataset_M2_5000 . . . 12
4.4.2 dataset_M1_10000 & dataset_M2_10000 . . . 12
4.4.3 dataset_M1_50000 & dataset_M2_50000 . . . 13
4.4.4 dataset_M1_100000 & dataset_M2_100000 . . . 14
4.4.5 dataset_M1_150000 & dataset_M2_150000 . . . 14
4.5 Conclusões Iniciais . . . 15

5 Benchmark **16**
5.1 Required Time . . . 16
5.2 Data Exchanged . . . 17

6 Conclusão **18**

1 Introdução

Este projeto têm como principal objetivo a exploração dos conceitos de Computação multipartidária segura com base nos protocolos PSI, Private set intersection protocols, bem como a sua comparação ao nível do tempo de execução e dados transferidos de acordo com o tamanho do dataset fornecido.

Estes protocolos permitem que duas entidades, cada uma com um conjunto de dados privado, calculem a interseção dos dados sem divulgar nenhuma informação sobre os mesmos. [1]

2 Problema

Iremos usar o PSI de forma a estudar e analisar a interceção de listas de pacientes de diversos médicos de um hospital. Este é importante devido à necessidade de privacidade que uma lista de pacientes de um médico exige pois esta informação é considerada confidencial e não deve ser exposta.

3 Datasets

Os Datasets contem informação relativa ao email, nome e idade dos pacientes. Tal que o email é constituido pelo nome próprio e apelido e foi gerado apartir de uma lista de 107 fornecedores de email. O nome próprio foi gerado apartir de uma lista de 2003 nomes e o apelido gerado apartir de uma lista de 3422 apelidos.

Os datasets intitulam-se de dataset_Mx_l tal que x é o número do médico e l o número total de linhas do dataset. Um exemplo do dataset pode se ver de seguinda:

```
email,name,age
Anne.Lampert@charter.net,Anne Lampert,31
Luis.Marquardt@hotmail.be,Luis Marquardt,66
Kordula.Schultz@live.com,Kordula Schultz,32
Fatima.Kempf@hotmail.com,Fatima Kempf,16
Wendelin.Rudolph@bellsouth.net,Wendelin Rudolph,40
Dennis.Gerner@safe-mail.net,Dennis Gerner,41
Armin.Locher@live.fr,Armin Locher,10
Reinhard.Klein@yahoo.com.mx,Reinhard Klein,68
```

Figura 1: Representação do dataset

Desta forma foram feitos os seguintes datasets:

- *dataset_M1_5000* & *dataset_M2_5000*
- *dataset_M1_10000* & *dataset_M2_10000*
- *dataset_M1_50000* & *dataset_M2_50000*
- *dataset_M1_100000* & *dataset_M2_100000*
- *dataset_M1_150000* & *dataset_M2_150000*

4 Protocolos

4.1 The naive hashing protocol

Nesta abordagem, cada *party* faz o hash do *input* e envia os respetivos hashes para as outras *parties*. Assim, cada party calcula a interseção com sua própria entrada de hash para encontrar as interseções.

Com a utilização deste protocolo, os médicos vão tomar conhecimento dos hashes dos outros médicos e comparar com os seus de forma a verificar quais os pacientes que têm em comum.

De forma a testar este protocolo será utilizado dois terminais, um para cada dataset usando os seguintes comandos:

- ./demo.exe -r 0 -p 0 -f projeto/dataset_M1_l.csv
- ./demo.exe -r 1 -p 0 -f projeto/dataset_M2_l.csv

4.1.1 dataset_M1_5000 & dataset_M2_5000

Os médicos que possuem 5000 pacientes apresentam 8 pacientes em comum.

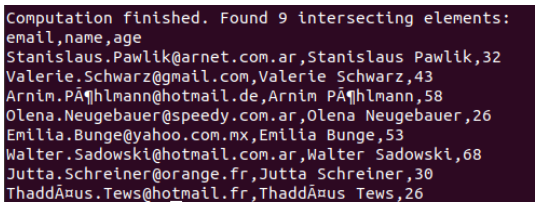

Figura 2: Execução do comando "./demo.exe -r 0 -p 0 -f projeto/dataset_M1_5000.csv".

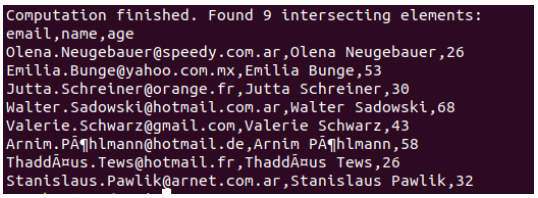

Figura 3: Execução do comando "./demo.exe -r 1 -p 0 -f projeto/dataset_M2_5000.csv".

4.1.2 dataset_M1_10000 & dataset_M2_10000

Os médicos que possuem 10000 pacientes apresentam 34 pacientes em comum.

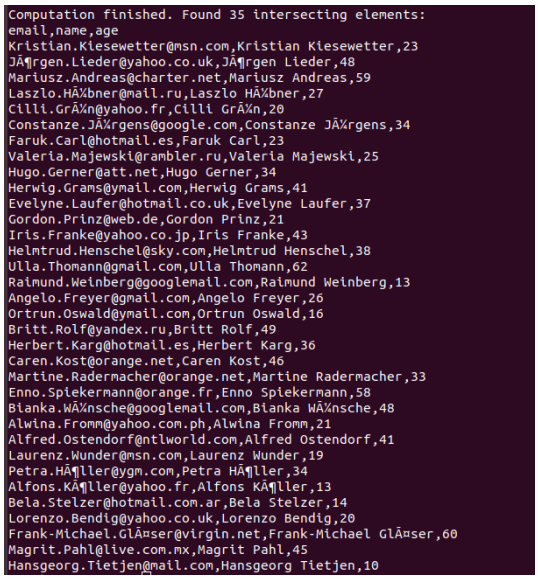

Figura 4: Execução do comando "./demo.exe -r 0 -p 0 -f projeto/dataset_M1_10000.csv".

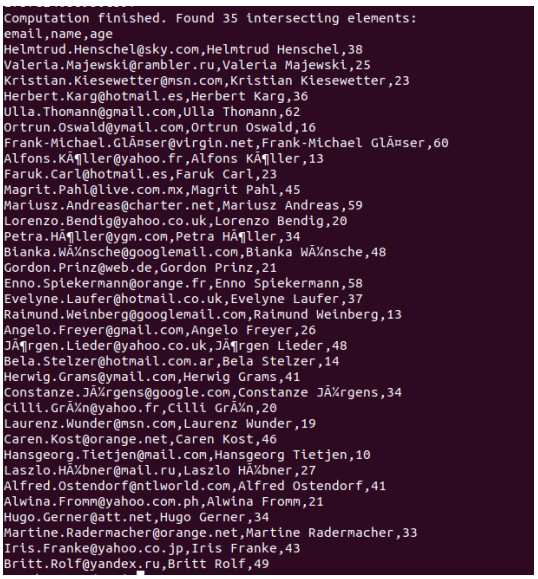

Figura 5: Execução do comando "./demo.exe -r 1 -p 0 -f projeto/dataset_M2_10000.csv".

4.1.3 dataset_M1_50000 & dataset_M2_50000

Os médicos que possuem 50000 pacientes apresentam 12 pacientes em comum.

```
Computation finished. Found 13 intersecting elements:
email,name,age
Gesine.Danner@bt.com,Gesine Danner,59
Dirk.Hansen@facebook.com,Dirk Hansen,30
Alwina.Lampe@gmx.net,Alwina Lampe,25
Klaus-Dieter.Hohmann@ygm.com,Klaus-Dieter Hohmann,37
Karl-Ludwig.BÃ¤hr@hotmail.co.uk,Karl-Ludwig BÃ¤hr,36
Silva.Rudolph@ymail.com,Silva Rudolph,26
Eveline.Jacobi@yandex.ru,Eveline Jacobi,52
Wolfhard.Pesch@zoho.com,Wolfhard Pesch,17
Hermine.Hirschfeld@orange.net,Hermine Hirschfeld,24
Auguste.Waldmann@facebook.com,Auguste Waldmann,18
SÃ¶nke.Schwarzer@fastmail.fm,SÃ¶nke Schwarzer,10
Kay-Uwe.Stiegler@hushmail.com,Kay-Uwe Stiegler,54
```

Figura 6: Execução do comando "./demo.exe -r 0 -p 0 -f projeto/dataset_M1_50000.csv".

```
Computation finished. Found 13 intersecting elements:
email,name,age
SÃ¶nke.Schwarzer@fastmail.fm,SÃ¶nke Schwarzer,10
Gesine.Danner@bt.com,Gesine Danner,59
Auguste.Waldmann@facebook.com,Auguste Waldmann,18
Silva.Rudolph@ymail.com,Silva Rudolph,26
Eveline.Jacobi@yandex.ru,Eveline Jacobi,52
Dirk.Hansen@facebook.com,Dirk Hansen,30
Kay-Uwe.Stiegler@hushmail.com,Kay-Uwe Stiegler,54
Hermine.Hirschfeld@orange.net,Hermine Hirschfeld,24
Klaus-Dieter.Hohmann@ygm.com,Klaus-Dieter Hohmann,37
Karl-Ludwig.BÃ¤hr@hotmail.co.uk,Karl-Ludwig BÃ¤hr,36
Wolfhard.Pesch@zoho.com,Wolfhard Pesch,17
Alwina.Lampe@gmx.net,Alwina Lampe,25
```

Figura 7: Execução do comando "./demo.exe -r 1 -p 0 -f projeto/dataset_M2_50000.csv".

4.1.4 dataset_M1_100000 & dataset_M2_100000

Os médicos que possuem 100000 pacientes apresentam 31 pacientes em comum. No entanto, é de notar que apresenta pacientes que não são comuns aos dois médicos como é o caso do paciente Heinz-Josef Graf que é paciente do médico 1 mas não do médico 2 e Carmen Kusch que é paciente do médico 2 mas não do médico 1.

```
Computation finished. Found 32 intersecting elements:
email,name,age
Heinz-Josef.Graf@att.net,Heinz-Josef Graf,15
Luka.Gebert@hotmail.com,Luka Gebert,39
Roderich.Trommer@ygm.com,Roderich Trommer,13
Aleksander.Hertel@yahoo.co.jp,Aleksander Hertel,27
Klaus-Dieter.Reusch@bt.com,Klaus-Dieter Reusch,42
Karl-Friedrich.Haller@ygm.com,Karl-Friedrich Haller,39
Heinz-Dieter.Seyfarth@qq.com,Heinz-Dieter Seyfarth,42
Denis.Born@hushmail.com,Denis Born,29
Hermann-Josef.Wollmann@laposte.net,Hermann-Josef Wollmann,49
Osman.Zimmerer@live.com.ar,Osman Zimmerer,26
Leonid.Achatz@fibertel.com.ar,Leonid Achatz,57
Ilona.Voss@safe-mail.net,Ilona Voss,23
Rebekka.Wagner@hush.com,Rebekka Wagner,26
Daniel.Philipp@gmx.com,Daniel Philipp,49
Rosalia.Gottfried@hotmail.co.uk,Rosalia Gottfried,36
Gerald.Brugger@prodigy.net.mx,Gerald Brugger,26
Moritz.Konrad@comcast.net,Moritz Konrad,46
Corinne.Braune@fibertel.com.ar,Corinne Braune,11
Ingrid.Obermeier@gmail.com,Ingrid Obermeier,47
Romuald.Gohlke@games.com,Romuald Gohlke,38
Corina.Kunzmann@nate.com,Corina Kunzmann,46
Ingbert.Tremmel@zoho.com,Ingbert Tremmel,60
Mohammad.Fehrenbach@pobox.com,Mohammad Fehrenbach,27
Mira.Jentzsch@yahoo.co.jp,Mira Jentzsch,28
Dorit.Basler@mail.ru,Dorit Basler,57
Jose.Spiekermann@orange.fr,Jose Spiekermann,66
Hans-Eberhard.Schnur@voo.be,Hans-Eberhard Schnur,39
Bayram.Greiner@inbox.com,Bayram Greiner,38
Eleni.Plum@verizon.net,Eleni Plum,27
GÃ¼nther.HÃ¶fling@hushmail.com,GÃ¼nther HÃ¶fling,51
Jovan.Brandstetter@wanadoo.co.uk,Jovan Brandstetter,13
```

Figura 8: Execução do comando "./demo.exe -r 0 -p 0 -f projeto/dataset_M1_100000.csv".

```
Computation finished. Found 32 intersecting elements:
email,name,age
Leonid.Achatz@fibertel.com.ar,Leonid Achatz,57
Rosalia.Gottfried@hotmail.co.uk,Rosalia Gottfried,36
Moritz.Konrad@comcast.net,Moritz Konrad,46
Luka.Gebert@hotmail.com,Luka Gebert,39
Mohammad.Fehrenbach@pobox.com,Mohammad Fehrenbach,27
GÃ¼nther.HÃ¶fling@hushmail.com,GÃ¼nther HÃ¶fling,51
Carmen.Kusch@sky.com,Carmen Kusch,30
Ingrid.Obermeier@gmail.com,Ingrid Obermeier,47
Karl-Friedrich.Haller@ygm.com,Karl-Friedrich Haller,39
Hans-Eberhard.Schnur@voo.be,Hans-Eberhard Schnur,39
Gerald.Brugger@prodigy.net.mx,Gerald Brugger,26
Eleni.Plum@verizon.net,Eleni Plum,27
Mira.Jentzsch@yahoo.co.jp,Mira Jentzsch,28
Osman.Zimmerer@live.com.ar,Osman Zimmerer,26
Rebekka.Wagner@hush.com,Rebekka Wagner,26
Bayram.Greiner@inbox.com,Bayram Greiner,38
Jose.Spiekermann@orange.fr,Jose Spiekermann,66
Hermann-Josef.Wollmann@laposte.net,Hermann-Josef Wollmann,49
Daniel.Philipp@gmx.com,Daniel Philipp,49
Dorit.Basler@mail.ru,Dorit Basler,57
Roderich.Trommer@ygm.com,Roderich Trommer,13
Jovan.Brandstetter@wanadoo.co.uk,Jovan Brandstetter,13
Romuald.Gohlke@games.com,Romuald Gohlke,38
Klaus-Dieter.Reusch@bt.com,Klaus-Dieter Reusch,42
Ingbert.Tremmel@zoho.com,Ingbert Tremmel,60
Aleksander.Hertel@yahoo.co.jp,Aleksander Hertel,27
Corina.Kunzmann@nate.com,Corina Kunzmann,46
Denis.Born@hushmail.com,Denis Born,29
Heinz-Dieter.Seyfarth@qq.com,Heinz-Dieter Seyfarth,42
Corinne.Braune@fibertel.com.ar,Corinne Braune,11
Ilona.Voss@safe-mail.net,Ilona Voss,23
```

Figura 9: Execução do comando "./demo.exe -r 1 -p 0 -f projeto/dataset_M2_100000.csv".

4.1.5 dataset_M1_150000 & dataset_M2_150000

Os médicos que possuem 10000 pacientes apresentam 29 pacientes em comum. No entanto, é de notar que apresenta pacientes que não são comuns aos dois médicos como é o caso, por exemplo, dos pacientes Baldur Tewes, Erhard Spiller e Sandra Gundlach que são pacientes do médico 1 mas não do médico 2. Ou dos pacientes Margaret Nitschke, Sibille Appelt e Evangelos Ammon que são pacientes do médico 2 mas não do médico 1.

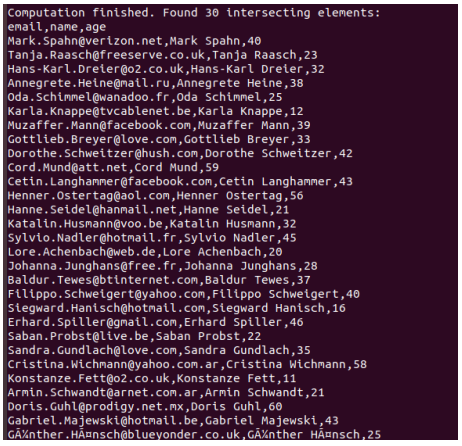

Figura 10: Execução do comando "./demo.exe -r 0 -p 0 -f projeto/dataset_M1_150000.csv".

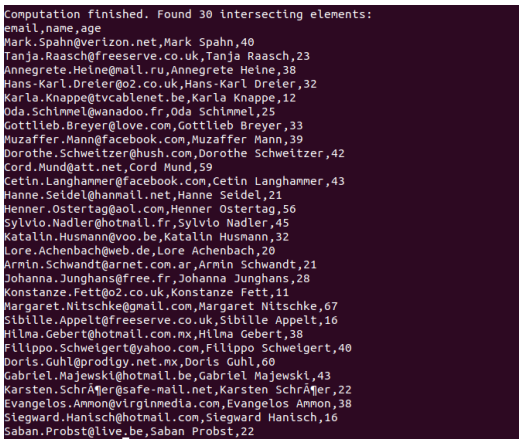

Figura 11: Execução do comando "./demo.exe -r 1 -p 0 -f projeto/dataset_M2_150000.csv".

4.2 The server-aided protocol

Este protocolo assume um terceiro que é responsável pelo cálculo da interseção. Desta forma, este protocolo é mais seguro comparativamente com o descrito anteriormente na medida que as *parties* nunca recebem a informação uma da outra. [2]

Com a utilização deste protocolo, os médicos enviam os hashes a um terceiro que os compara de forma a verificar quais os pacientes que os médicos têm em comum e envia-lhes as intercepões.

De forma a testar este protocolo será utilizado três terminais, um para cada o servidor e dois para os datasets, usando os seguintes comandos:

- ./demo.exe -r 0 -p 1 -f README.md -a 127.0.0.1
- ./demo.exe -r 1 -p 1 -f projeto/dataset_M1_l.csv 127.0.0.1
- ./demo.exe -r 1 -p 1 -f projeto/dataset_M2_l.csv 127.0.0.1

4.2.1 dataset_M1_5000 & dataset_M2_5000

Á semelhança do protocolo anteriormente, foram encontradas 8 pacientes em comum.

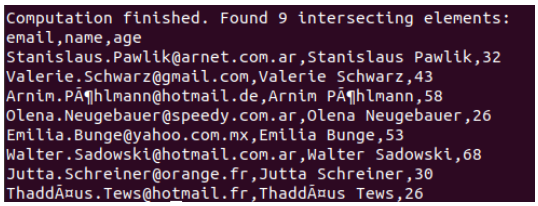

Figura 12: Execução do comando "./demo.exe -r 1 -p 1 -f projeto/dataset_M1_5000.csv 127.0.0.1".

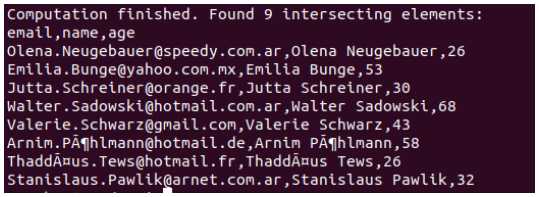

Figura 13: Execução do comando "./demo.exe -r 1 -p 1 -f projeto/dataset_M2_5000.csv 127.0.0.1".

4.2.2 dataset_M1_10000 & dataset_M2_10000

Á semelhança do protocolo anteriormente, foram encontradas 34 pacientes em comum.

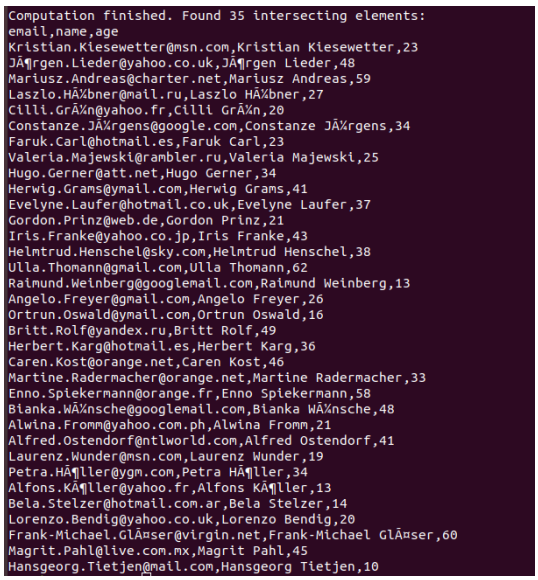

Figura 14: Execução do comando "./demo.exe -r 1 -p 1 -f projeto/dataset_M1_10000.csv 127.0.0.1".

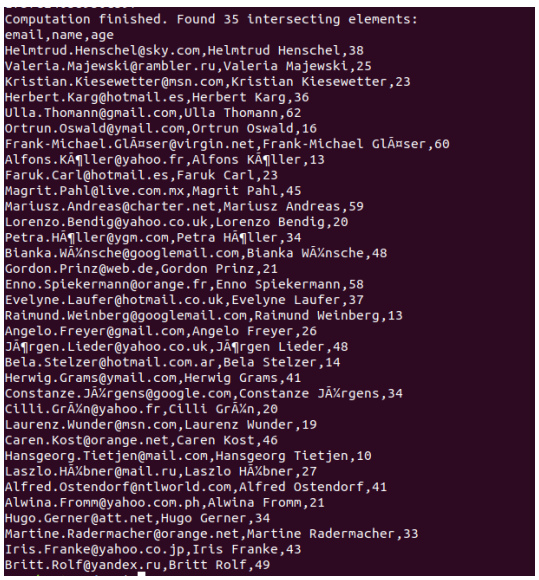

Figura 15: Execução do comando "./demo.exe -r 1 -p 1 -f projeto/dataset_M2_10000.csv 127.0.0.1".

4.2.3 dataset_M1_50000 & dataset_M2_50000

Á semelhança do protocolo anteriormente, foram encontradas 12 pacientes em comum.

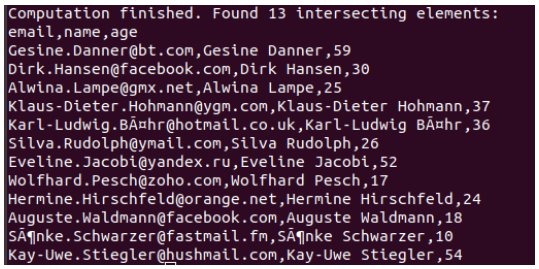

Figura 16: Execução do comando "./demo.exe -r 1 -p 1 -f projeto/dataset_M1_50000.csv 127.0.0.1".

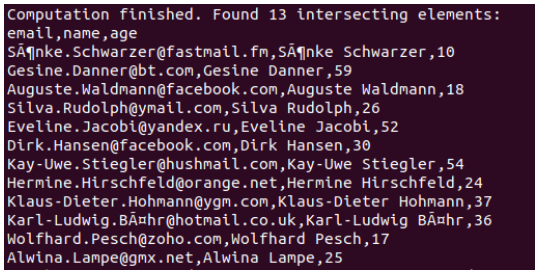

Figura 17: Execução do comando "./demo.exe -r 1 -p 1 -f projeto/dataset_M2_50000.csv 127.0.0.1".

4.2.4 dataset_M1_100000 & dataset_M2_100000

Ao contrário do protocolo anterior, foram agora encontrados apenas 30 pacientes em comum e como podemos verificar a intercepção que desapareceu foi a que continha os falsos positivos.

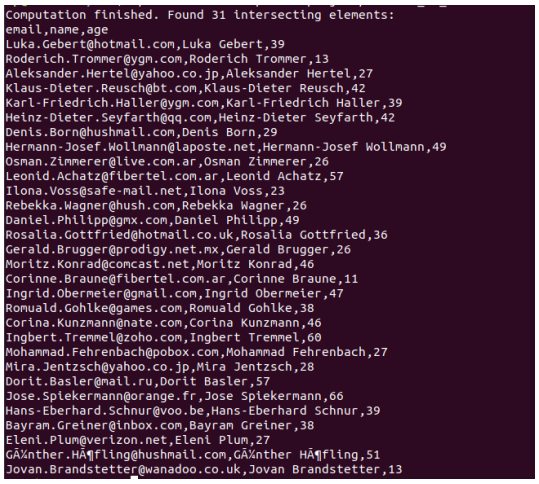

Figura 18: Execução do comando "./demo.exe -r 1 -p 1 -f projeto/dataset_M1_100000.csv 127.0.0.1".

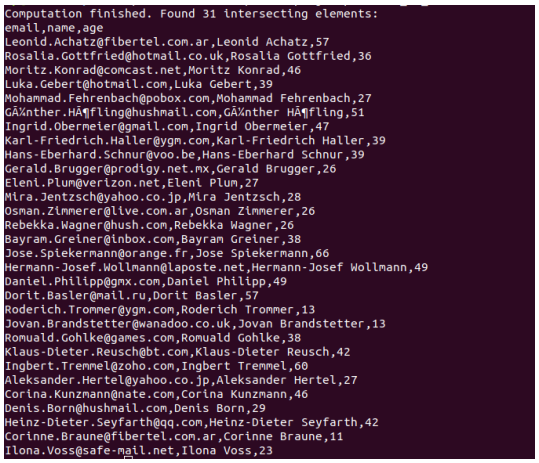

Figura 19: Execução do comando "./demo.exe -r 1 -p 1 -f projeto/dataset_M2_100000.csv 127.0.0.1".

4.2.5 dataset_M1_150000 & dataset_M2_150000

Ao contrário do protocolo anterior, foram agora encontrados apenas 24 pacientes em comum e como podemos verificar a intercepção que desapareceu foi a que continha os falsos positivos.

```
Computation finished. Found 25 intersecting elements:
email,name,age
Mark.Spahn@verizon.net,Mark Spahn,40
Tanja.Raasch@freeserve.co.uk,Tanja Raasch,23
Hans-Karl.Dreier@o2.co.uk,Hans-Karl Dreier,32
Annegrete.Heine@mail.ru,Annegrete Heine,38
Oda.Schimmel@wanadoo.fr,Oda Schimmel,25
Karla.Knappe@tvcablenet.be,Karla Knappe,12
Muzaffer.Mann@facebook.com,Muzaffer Mann,39
Gottlieb.Breyer@love.com,Gottlieb Breyer,33
Dorothe.Schweitzer@hush.com,Dorothe Schweitzer,42
Cord.Mund@att.net,Cord Mund,59
Cetin.Langhammer@facebook.com,Cetin Langhammer,43
Henner.Ostertag@aol.com,Henner Ostertag,56
Hanne.Seidel@hanmail.net,Hanne Seidel,21
Katalin.Husmann@voo.be,Katalin Husmann,32
Sylvio.Nadler@hotmail.fr,Sylvio Nadler,45
Lore.Achenbach@web.de,Lore Achenbach,20
Johanna.Junghans@free.fr,Johanna Junghans,28
Filippo.Schweigert@yahoo.com,Filippo Schweigert,40
Siegward.Hanisch@hotmail.com,Siegward Hanisch,16
Saban.Probst@live.be,Saban Probst,22
Konstanze.Fett@o2.co.uk,Konstanze Fett,11
Armin.Schwandt@arnet.com.ar,Armin Schwandt,21
Doris.Guhl@prodigy.net.mx,Doris Guhl,60
Gabriel.Majewski@hotmail.be,Gabriel Majewski,43
```

Figura 20: Execução do comando "./demo.exe -r 1 -p 1 -f projeto/dataset_M1_150000.csv 127.0.0.1".

```
Computation finished. Found 25 intersecting elements:
email,name,age
Mark.Spahn@verizon.net,Mark Spahn,40
Tanja.Raasch@freeserve.co.uk,Tanja Raasch,23
Annegrete.Heine@mail.ru,Annegrete Heine,38
Hans-Karl.Dreier@o2.co.uk,Hans-Karl Dreier,32
Karla.Knappe@tvcablenet.be,Karla Knappe,12
Oda.Schimmel@wanadoo.fr,Oda Schimmel,25
Gottlieb.Breyer@love.com,Gottlieb Breyer,33
Muzaffer.Mann@facebook.com,Muzaffer Mann,39
Dorothe.Schweitzer@hush.com,Dorothe Schweitzer,42
Cord.Mund@att.net,Cord Mund,59
Cetin.Langhammer@facebook.com,Cetin Langhammer,43
Hanne.Seidel@hanmail.net,Hanne Seidel,21
Henner.Ostertag@aol.com,Henner Ostertag,56
Sylvio.Nadler@hotmail.fr,Sylvio Nadler,45
Katalin.Husmann@voo.be,Katalin Husmann,32
Lore.Achenbach@web.de,Lore Achenbach,20
Armin.Schwandt@arnet.com.ar,Armin Schwandt,21
Johanna.Junghans@free.fr,Johanna Junghans,28
Konstanze.Fett@o2.co.uk,Konstanze Fett,11
Filippo.Schweigert@yahoo.com,Filippo Schweigert,40
Doris.Guhl@prodigy.net.mx,Doris Guhl,60
Gabriel.Majewski@hotmail.be,Gabriel Majewski,43
Siegward.Hanisch@hotmail.com,Siegward Hanisch,16
Saban.Probst@live.be,Saban Probst,22
```

Figura 21: Execução do comando "./demo.exe -r 1 -p 1 -f projeto/dataset_M2_150000.csv 127.0.0.1".

4.3 The Diffie-Hellman-based PSI protocol

O protocolo é baseado na troca de chaves de Diffie-Hellman, que permite que duas partes estabeleçam uma chave secreta compartilhada num canal de comunicação inseguro. Desta forma cada parte gera um par de chaves privada de Diffie-Hellman e publica a sua chave pública, calculando a interseção sem revelar os elementos reais dos conjuntos entre si. [3]

Desta forma os médicos geram uma chave chave publica e uma chave privada de Diffie-Hellman e publicam a sua chave pública, depois de enviar as suas listas encriptadas com uma função hash usam a chave privada para desencriptar e calcular a interseção.

De forma a testar este protocolo será utilizado dois terminais, um para cada dataset usando os seguintes comandos:

- ./demo.exe -r 0 -p 2 -f projeto/dataset_M1_l.csv
- ./demo.exe -r 1 -p 2 -f projeto/dataset_M2_l.csv

4.3.1 dataset_M1_5000 & dataset_M2_5000

À semelhança do ocorrido nos protocolos anteriores foram encontrados 8 pacientes em comum entre os dois médicos.

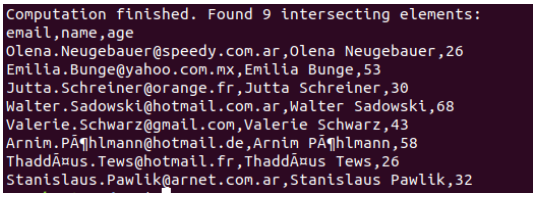

Figura 22: Execução do comando "./demo.exe -r 1 -p 2 -f projeto/dataset_M2_5000.csv".

4.3.2 dataset_M1_10000 & dataset_M2_10000

À semelhança do ocorrido nos protocolos anteriores foram encontrados 34 pacientes em comum entre os dois médicos.

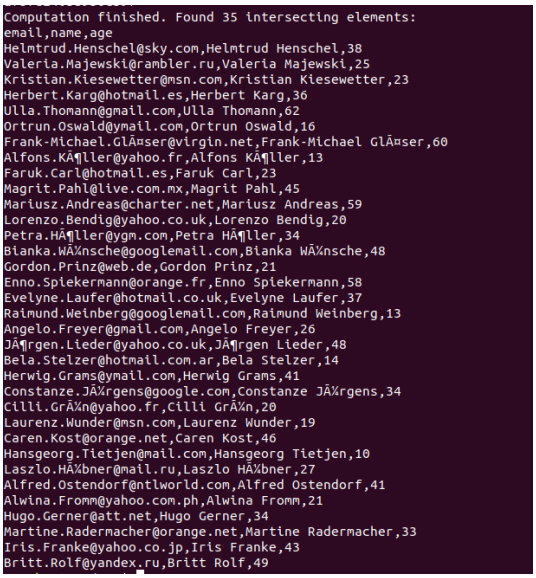

Figura 23: Execução do comando "./demo.exe -r 1 -p 2 -f projeto/dataset_M2_10000.csv".

4.3.3 dataset_M1_50000 & dataset_M2_50000

Ao contrario dos protocolos anteriores, foi detetado 15 pacientes em comum entre os dois médicos.

No entanto, tendo em conta que este protocolo só demonstra as interceções de um médico fomos confirmar diretamente ao dataset onde, rapidamente, chegamos à conclusão que as interceções "extras"tratam se de falsos positivos.

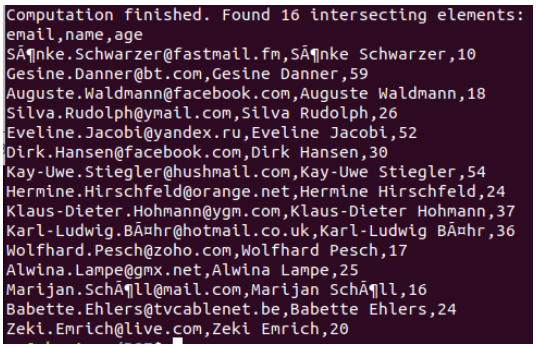

Figura 24: Execução do comando "./demo.exe -r 1 -p 2 -f projeto/dataset_M2_50000.csv".

4.3.4 dataset_M1_100000 & dataset_M2_100000

Á semelhança do *The naive hashing protocol* foram encontrados 31 pacientes em comum. No entanto comparando os outputs podemos verificar que o falso positivo não se trata da paciente Carmen Kusch.

Assim, comparando com o protocolo *The server-aided protocol* que não apresentou falsos positivos podemos verificar que o falso positivo é o paciente Ryszard Haag

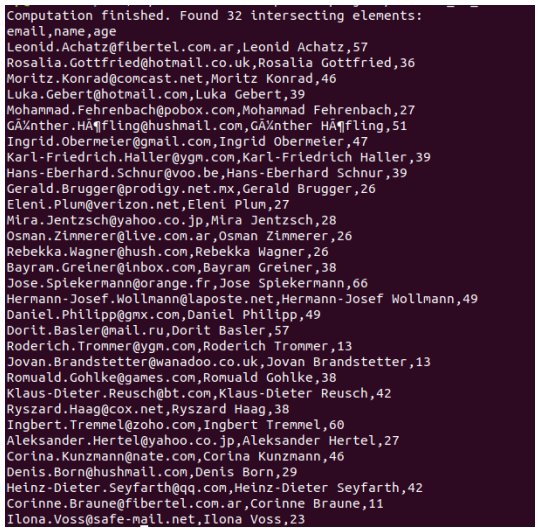

Figura 25: Execução do comando "./demo.exe -r 1 -p 2 -f projeto/dataset_M2_100000.csv".

4.3.5 dataset_M1_150000 & dataset_M2_150000

Tendo em conta que novamente foram detetadas mais pacientes em comum do que o suposto formos comparar com os outputs dos protocolos referidos anteriormente.

Assim podemos verificar que os pacientes Karl-Werner Schuh, Stefania Jost e Cora Mai são falsos positivos.

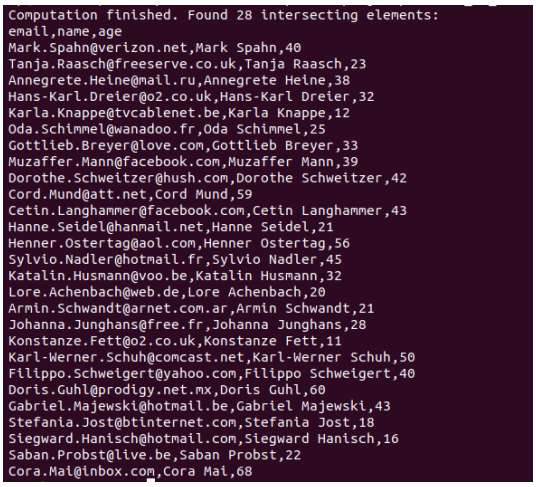

Figura 26: Execução do comando "./demo.exe -r 1 -p 2 -f projeto/dataset_M2_150000.csv".

4.4 The OT-based PSI protocol

Este protocolo baseia-se no conceito de Transferência Oblíqua (Oblivious Transfer - OT), que permite que um remetente transfira uma parte da informação para um receptor, sem que o remetente saiba que parte foi escolhida pelo receptor e sem que o receptor saiba sobre a restante informação. Assim a informação é transmitida de forma segura, preservando a privacidade. [4]

Desta forma um médico irá transferir parte da sua lista de pacientes para o outro médico usando o protocolo OT para realizar transferências oblíquas. Enquanto o outro médico irá receber essa informação e parte da informação da sua própria lista. Usando as informações recebidas através do OT, os médicos são capazes de determinar quais pacientes estão presentes em ambas as listas sem revelar informações sobre os pacientes que não estão na interseção.

De forma a testar este protocolo será utilizado dois terminais, um para cada dataset usando os seguintes comandos:

- ./demo.exe -r 0 -p 3 -f projeto/dataset_M1_1.csv
- ./demo.exe -r 1 -p 3 -f projeto/dataset_M2_1.csv

4.4.1 dataset_M1_5000 & dataset_M2_5000

À semelhança dos protocolos utilizados anteriormente existem 8 pacientes em comum entre os médicos.

Para alem disso, pelo output é possivel afirmar que este protocolo utilizou 6002 OTs (Oblivious Transfers) de forma a permitir a transferência de informações de forma segura, enquanto o protocolo está definido para ter 60 de comprimento em Bits do elemento e 72 de comprimento em Bits da máscara.

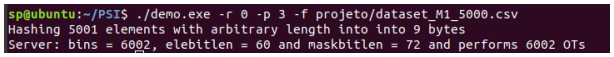

Figura 27: Execução do comando "./demo.exe -r 0 -p 3 -f projeto/dataset_M1_5000.csv".

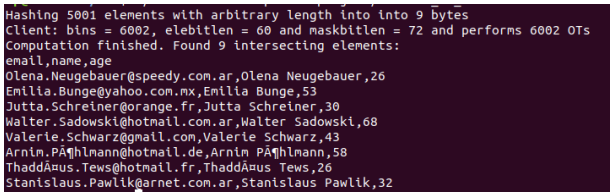

Figura 28: Execução do comando "./demo.exe -r 1 -p 3 -f projeto/dataset_M2_5000.csv".

4.4.2 dataset_M1_10000 & dataset_M2_10000

À semelhança dos protocolos utilizados anteriormente existem 34 pacientes em comum entre os médicos.

Para alem disso, pelo output é possivel afirmar que este protocolo utilizou 12002 OTs (Oblivious Transfers) de forma a permitir a transferência de informações de forma segura, enquanto o protocolo está definido para ter 59 de comprimento em Bits do elemento e 72 de comprimento em Bits da máscara.

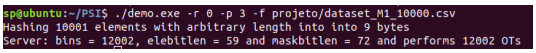

Figura 29: Execução do comando "./demo.exe -r 0 -p 3 -f projeto/dataset_M1_10000.csv".

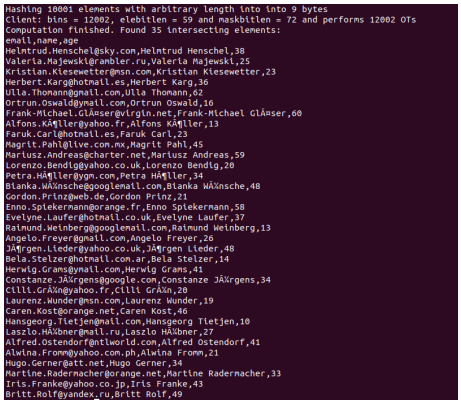

Figura 30: Execução do comando "./demo.exe -r 1 -p 3 -f projeto/dataset_M2_10000.csv".

4.4.3 dataset_M1_50000 & dataset_M2_50000

À semelhança dos protocolos utilizados anteriormente existem 12 pacientes em comum entre os médicos.

Para alem disso, pelo output é possivel afirmar que este protocolo utilizou 60002 OTs (Oblivious Transfers) de forma a permitir a transferência de informações de forma segura, enquanto o protocolo está definido para ter 57 de comprimento em Bits do elemento e 72 de comprimento em Bits da máscara.

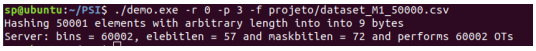

Figura 31: Execução do comando "./demo.exe -r 0 -p 3 -f projeto/dataset_M1_50000.csv".

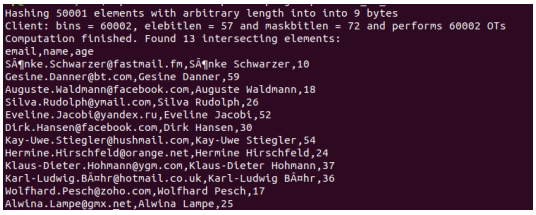

Figura 32: Execução do comando "./demo.exe -r 1 -p 3 -f projeto/dataset_M2_50000.csv".

4.4.4 dataset_M1_100000 & dataset_M2_100000

À semelhança dos protocolos utilizados anteriormente existem 30 pacientes em comum entre os médicos.

Para alem disso, pelo output é possivel afirmar que este protocolo utilizou 120002 OTs (Oblivious Transfers) de forma a permitir a transferência de informações de forma segura, enquanto o protocolo está definido para ter 64 de comprimento em Bits do elemento e 80 de comprimento em Bits da máscara.

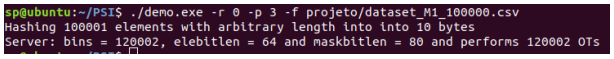

Figura 33: Execução do comando "./demo.exe -r 0 -p 3 -f projeto/dataset_M1_100000.csv".

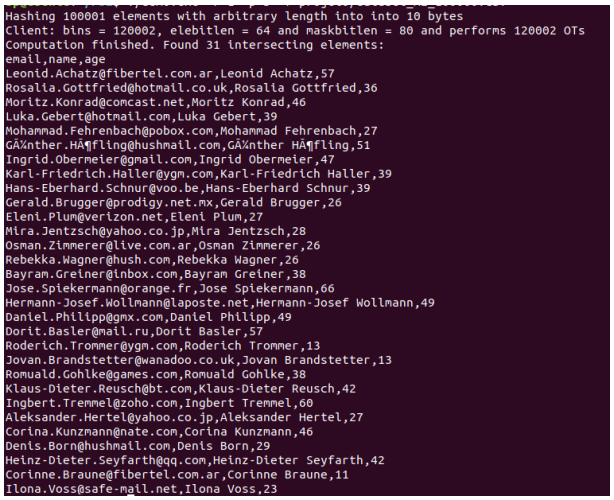

Figura 34: Execução do comando "./demo.exe -r 1 -p 3 -f projeto/dataset_M2_100000.csv".

4.4.5 dataset_M1_150000 & dataset_M2_150000

À semelhança dos protocolos utilizados anteriormente existem 24 pacientes em comum entre os médicos.

Para alem disso, pelo output é possivel afirmar que este protocolo utilizou 180002 OTs (Oblivious Transfers) de forma a permitir a transferência de informações de forma segura, enquanto o protocolo está definido para ter 63 de comprimento em Bits do elemento e 80 de comprimento em Bits da máscara.

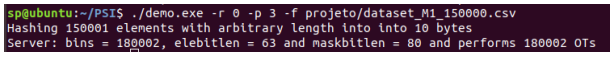

Figura 35: Execução do comando "./demo.exe -r 0 -p 3 -f projeto/dataset_M1_150000.csv".

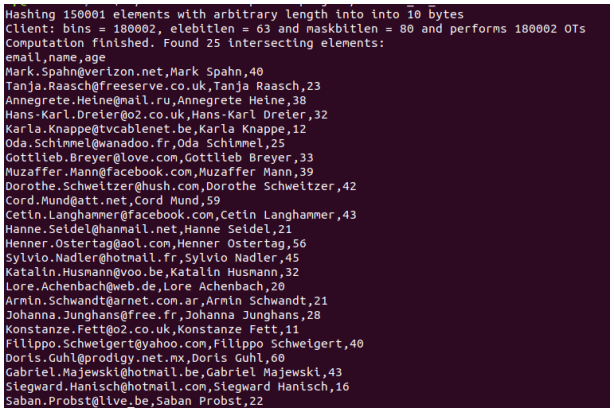

Figura 36: Execução do comando "./demo.exe -r 1 -p 3 -f projeto/dataset_M2_150000.csv".

4.5 Conclusões Iniciais

Durante o teste dos diferentes protocolos foi possivel notar o seguinte:

- O protocolo *The server-aided protocol* e *The OT-based PSI protocol* não apresentaram falsos negativos ao contrário dos restantes protocolos. Isto deve-se ao facto destes protocolos serem projetados de forma a minimizar ou eliminar a ocorrência de falsos positivos.

  No caso do *The server-aided protocol* como apenas o servidor tem conhecimento das duas listas de dados e executa diretamente as comparações, pode garantir que apenas os elementos comuns às duas listas sejam considerados na interseção. Reduzindo a ocorrência de falsos positivos, já que a lógica de comparação é controlada diretamente pelo servidor.

  Já no OT-Based PSI Protocol como usa transferências oblíquas permite que o receptor receba informações sobre os elementos da sua própria lista, sem que o emissor saiba quais elementos foram selecionados. Dessa forma, apenas os elementos presentes em ambas as listas são transferidos, eliminando a ocorrência de falsos positivos. E, devido à segura mantida pelas transferências oblíquas, o receptor não consegue retirar informações sobre os elementos exclusivos da lista do emissor.

- Nos protocolos *The naive hashing protocol* e *The Diffie-Hellman-based PSI* é possivel observar que com o aumento de linhas do dataset há uma maior quantidade de falsos positivos advendo-se das limitações destes protocolos.

  Como o protocolo *Naive Hashing Protocol* apenas cria um conjunto de valores encriptados, sem garantir a exclusividade dos valores hash pode levar a colisões de hash (ou seja, diferentes informações com o mesmo valor hash), tal que com o aumento das linhas de um dataset a probabilidade disto acontecer aumenta, levando a uma maior quantidade de falsos positivos.

  Já no *Diffie-Hellman-based PSI Protocol*, com o aumento do tamanho do dataset, a geração de chaves e os cálculos associados ao protocolo *Diffie-Hellman* tornam-se mais complexos e propensos a erros, o que pode resultar em falsos positivos.

- Para alem disso, os protocolos *The Diffie-Hellman-based PSI protocol* e *The OT-based PSI protocol* têm como desvantagem não permitir uma comparação direta das interceções, havendo uma maior necessidade de comparar o resultado com o resultado de outros protocolos e/ou confirmar os resultados com recurso aos proprios datasets.

5 Benchmark

De forma a analisar e avaliar os protocolos de Private Set Intersection executamos o *psi.exe* em dois terminais. Este necessita da introdução do protocolo (p) a ser estudado, no entanto não necessita da introdução de nenhum dataset uma vez que cria o seu próprio input dado um determinado número de linhas (l) .

- ./psi.exe -r 0 -p p -b 16 -n l
- ./psi.exe -r 1 -p p -b 16 -n l

É de notar que uma vez que o protocolo *The server-aided protocol* têm um bug para o benchmark apenas serão testados os restantes protocolos.

No final de cada execução é dado o tempo necessário em segundos (required time), os dados enviados (data sent) e os dados recebidos em MegaBytes(data received), sendo que a sua soma deve ser considerada como os dados trocados (data exchanged). Na figura 37 podemos observar o resultado das diversas execuções.

The Naive Hashing Protocol		
	required time	data exchanged
5000	0,1	0,00
10000	0,2	0,2
50000	1,3	0,8
100000	2,7	1
150000	3,6	2,8
200000	4,6	3,8
250000	6	4,8
300000	7,4	5,8
350000	8,9	6,6
400000	9,4	7,6
450000	11,8	8,6

The Diffie-Hellman-based PSI protocol		
	required time	data exchanged
5000	5,1	0,5
10000	9,7	1,1
50000	50,1	5,1
100000	98,8	10,1
150000	156,1	15,2
200000	206,6	20,3
250000	268,5	25,3
300000	308	30,3
350000	358,2	35,4
400000	425,7	40,4
450000	470,8	45,5

The OT-based PSI protocol		
	required time	data exchanged
5000	0,4	0,5
10000	0,5	1,1
50000	0,6	5
100000	0,8	10,2
150000	1,1	15,3
200000	1,3	18,4
250000	1,5	25,5
300000	1,8	30,6
350000	2,1	35,7
400000	2,3	40,8
450000	3,1	45,9

Figura 37: Resultados do benchmark

Perante os resultados do benchmark criamos gráficos representativos do tempo necessário em função do aumento de linhas num dataset (figura 38) e os dados enviados em função do aumento de linhas num dataset (figura 40).

5.1 Required Time

Observando a figura 38 podemos afirmar que o *Diffie-Hellman-based PSI protocol* necessita relativamente de um maior tempo comparativamente aos restantes protocolos. Como referido anteriormente, com o aumento do tamanho do dataset, a geração de chaves e os cálculos associados ao protocolo *Diffie-Hellman* tornam-se mais complexo. Assim para alem de um maior numero de falsos positivos podemos tambem concluir que o tempo de execução aumenta conforme o aumento de linhas.

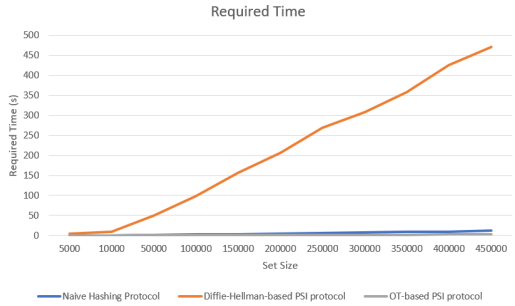

Figura 38: Representação do tempo necessário em função do tamanho do dataset

De forma a retirarmos melhores conclusões relativamente aos protocolos *Naive hashing protocol* e *OT-based PSI protocol* foi feito um gráfico adicional. Desta forma é possivel observar que apesar do protocolo *OT-based PSI protocol* apresentar valores iniciais ligeiramente mais altos que o protocolo *Naive hashing protocol*, este rápidamente ultrapassa. Isto é justificado pela necessidade de no protocolo *Naive hashing protocol* ser enviada toda a informação dos datasets enquanto no protocolo *OT-based PSI protocol* apenas é enviado parte da informação contribuindo para a eficiencia deste protocolo.

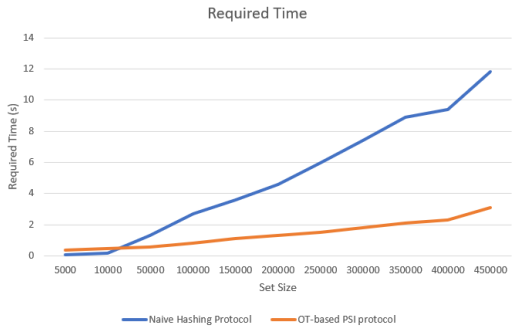

Figura 39: Representação do tempo necessário em função do aumento do tamanho do dataset em maior detalhe

5.2 Data Exchanged

Observando a figura 40 podemos afirmar que o protocolo *Naive hashing protocol* troca relativamente menos informação que os protocolos *Diffie-Hellman-based PSI protocol* e *OT-based PSI protocol*.

Isto é justificado pelo facto de no protocolo *Naive hashing protocol* a informação trocada é relativamente menor uma vez que cada *party* gera os hashes dos seus próprios conjuntos de dados e envia esses hashes para a outra *party*. Estes hashes, como apenas são utilizados para identificar a correspondência entre os elementos dos conjuntos de dados, não contêm informações detalhadas sobre os elementos em si. Portanto, a quantidade de informação trocada é limitada aos hashes gerados.

Já nos protocolos *Diffie-Hellman-based PSI protocol* e *OT-based PSI protocol* a troca de informações é maior uma vez que envolve a troca de chaves públicas e a realização de cálculos criptográficos para determinar as correspondências, implicando numa maior quantidade de informações a ser trocada entre as *parties* para permitir a realização dos cálculos enquanto garante a privacidade dos dados.

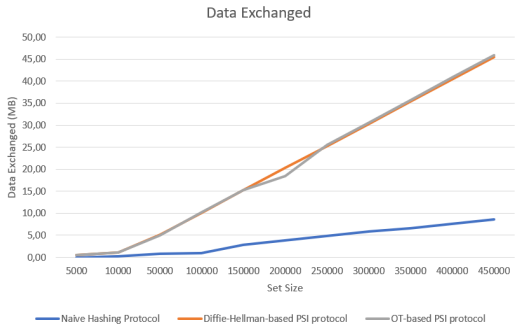

Figura 40: Representação dos dados enviados em função do aumento de linhas num dataset

6 Conclusão

Este projeto contribuiu para o entendimento do conceito de Private Set Intersection (PSI) e dos protocolos existentes. Foi realizado um estudo sobre a ocorrência de falsos positivos, o tempo de execução e a informação transmitida, e chegamos à conclusão de que, para conjuntos de dados grandes, o uso do protocolo *Diffie-Hellman-based PSI protocol* não é recomendado devido à alta probabilidade de falsos positivos e à necessidade de um tempo de execução mais longo. O protocolo Naive Hashing, apesar de ter um tempo de execução menor em comparação com o *Diffie-Hellman-based PSI protocol*, também apresentou uma grande quantidade de falsos positivos, e sua utilização não é aconselhada devido à baixa segurança que oferece.

Além disso, embora não tenha sido realizado um benchmark do protocolo *The Server-Aided Protocol*, este pode ser uma boa opção para grandes conjuntos de dados, pois mantém a segurança ao permitir a análise das interseções por meio de um servidor, evitando a ocorrência de falsos positivos. Da mesma forma, o uso do protocolo *OT-based PSI protocol* não demonstrou a presença de falsos positivos e requer pouco tempo de execução.

Referências

[1] Decentralized Thoughts. The Private Set Intersection (PSI) Protocol of the Apple CSAM Detection System. `https://decentralizedthoughts.github.io/2021-08-29-the-private-set-intersection-psi-protocol-of-the-apple-csam-detection-system/`, agosto 2021.

[2] Seny Kamara, Payman Mohassel, Mariana Raykova, and Shaza Sadeghian. Scaling private set intersection to billion-element sets. In *Financial Cryptography and Data Security (FC'14)*, Lecture Notes in Computer Science (LNCS). Springer, 2014.

[3] Catherine Meadows. A more efficient cryptographic matchmaking protocol for use in the absence of a continuously available third party. In *IEEE Symposium on Security and Privacy (SP'86)*, pages 134–137. IEEE, 1986.

[4] Interseção de conjunto privado escalável com base na extensão ot. Cryptology ePrint Archive, Paper 2016/930.