Example Paper for the Studierendentagung 2018 – How it Should Look Like –

N. Jobst [1], and D. Musterprofessor [5]

[1] Medizinische Informatik, Universität zu Lübeck, niklas.jobst@student.uni-luebeck.de
[2] Institute of Wisdom, Universität zu Lübeck, musterprofessor@wisdom.uni-luebeck.de

Abstract

The goal of this practicum was to discuss, how two parties can compute the similarity of their DNA, without allowing any of them to gain knowledge about the genetic code of the other one.
The basis for these computations had been already existing methods, with which one can compute the intersection of two Sets privacy preserving. For this practicum I have implemented two of them in Java. They both rely Bloom Filters as well as on a cryptosystem, either Elgamal or Paillier.
While both methods are able to compute the DNA Set-Intersection in an agreeable time, the method based on Elgamal has been significantly faster due to its faster encryption.

1 Introduction

There are currently

2 Material and Methods

I have implemented two methods. B Both are using therefore

2.1 Bloom Filter

Bloom Filters are a technique to test whether specific data elements are part of a dataset or not. They consist out of an with zeros initialized m bit long array and k hash functions which are mapped on the array.
To initialize a bloom filter, all hash functions are used on every data element of the dataset of interest. The positions in the array, which are equal to the outcome of the hash functions are set to one.
To test whether a element is part of the dataset, all hash functions are used on the element.
If all positions of the outcome are set to one in the array one can assume that the element is part of the dataset, through bloom filters are not fully resistant to false positives.

2.2 Elgamal

The public-key cryptosystem Elgamal is a enhancement of the Diffie-Hellmann key exchange.
For key generation the client choses a cyclic group Z of the order q with the generator g. The client then choses a random number $a < q$ as the secret key. Next the client computes $P = g^{\mathrm{a}}$, which is used together with g, q, Z as the public key.

To encrypt a message $m \in Z_{\mathrm{q}}$, the server then choses a random number $r < q$. The server then computes $c_1 = g^{\mathrm{r}}$ as well as $c_2 = P^{\mathrm{r}} * m$. The cipher text consists the out of $C = (c_1, c_2)$.

The client then can compute $\Sigma = c_1^{-q} * c_2$ to decrypt the ciphertext.

Elgamal is homomorph for multiplication

$$E(m_1 * m_2) = (E(m_1) * E(m_2))$$

2.3 Paillier

For key generation the client choses to prime numbers p, q with $ggt(pq, (p-1)(q-1)) = 1undn = pq$. The generator g is then chosen, so that $g \in (\mathbb{Z}n^2\mathbb{Z})$ and$n$ divides the order of g. The public key consists then out of(n, g). The client computes $\lambda = kgV(p-1, q-1)$ as the secret key.

To encrypt a message $m \in \mathbb{Z}_n^*$ the server choses a random number $r \in \mathbb{Z}_{n^2}^*$. The server computes the cipher text $c = g^m * r^n \mod n^2$.

For decryption the client needs first $L(x) = \frac{(x-1)}{n}$. Then he can compute the plain text $m = \frac{L(c^\lambda \ mod \ n^2)}{L(g^\lambda \ mod \ n^2)} \ mod \ n$.

Elgamal is homomorph for addition

$$E(m_1 + m_2) = (E(m_1) * E(m_2))$$

2.4 Algorithm based on Elgamal

The method I am going to present here was first published in ....
The client first constructs a bloomfilter over his data. Using the elgamal cryptosystem, the client then encrypts every single bit of the bloomfilter array. For this encryption the message for each bit is constructed in the following way: $m[i] = g^B F_C lient[i]$. By doing so, the message $m_[i]$ has the value one coded in case the bloomfilter has on the

$$S_\mathrm{i} = pk^{r_i} * \begin{cases} g^0 = 1 \; bei \; BF_1[i] = 1 \\ g^1 = g \; bei \; BF_1[i] = 0 \end{cases}$$

The ciphertext the client constructs has the following form:

$$(R_i, S_i) = (g^{\mathrm{r_i}}, pk^{\mathrm{r_i}} * g^{\text{1-BF}_1\text{[i]}})$$

The client now sends the ciphertext alongside with the public key and the Bloomfilter parameters to the Server.
In the second step the Server now creates a Bloomfilter over his data, using the the Bloomfilter parameters of the client. Next the server selects all positions in which his Bloomfilter has a zero entry. He then identifies the corresponding entries in the ciphertext of the client and multiplies them together. The results are then rerandomised by the server.

$$V = (g^\mathrm{s} * \Pi_{i:BF_2[i]=0} R_i)$$

$$W = (pk^\mathrm{s} * \Pi_{i:BF_2[i]=0} S_i)$$

Finally the server sends V and W back to the client.
In the last step, the client decrypts V and W by using his private key.

$$\Sigma = W * V^{-sk}$$

Due to the fact, that $pk = g^{sk}$, the Equation can be displayed in the following way:

$$\Sigma = (g^{sk*s+r_{i_1}+r_{i_2}+ \; .. \; +r_{i_k}} * g^{-sk*s+r_{i_1}+r_{i_2}+ \; .. \; +r_{i_k}} * g^z)$$

After canceling the equation can be shortened to:

$$\Sigma = g^z$$

z equals here the number number of positions where the client as well as the server have an zero entry in the Bloomfilters.
The client now can approximate the number of elements, which just one of the participats owns:

$$|X| = \frac{ln(\frac{z}{m})}{k * \; ln(1 - \frac{1}{m})}$$

2.5 Algorithm based on Paillier

The method I am going to present here was first published in ..
The client first constructs a Bloomfilter over his data. Using the elgamal cryptosystem, the client then encrypts every single bit of the Bloomfilter array.

$$c_i = (g^{\mathrm{IBF[i]}} * {r_i}^\mathrm{n})$$

$$C_\mathrm{i} = r_i^n \; * \begin{cases} g^0 = 1 \; bei \; BF_1[i] = 1 \\ g^1 = g \; bei \; BF_1[i] = 0 \end{cases}$$

pk: public key, sk: private key, g: Generator r_i: Zufallszahlen aus Z_q
The server now creates a bloomfilter for every element of his dataset. Then he mulitplicates the clients ciphertext on those positions, where $BF_{server[j]} = 1$:

$$V_j = (g^{IBF_{i_1}+IBF_{i_2}+ \; ... \; +IBF_{i_k}} \; * \; r_{i_1}^n \; * \; r_{i_2}^n \; * \; ... \; * \; r_{i_k}^n)$$

$$V_j = \; r_{i_1}^n * r_{i_2}^n * ... * r_{i_k}^n \begin{cases} g^{1+1+1+ \; ... \; +1} \; wenn \; BF_c = 0, BF_{s[j]} = 1 \\ g^{0+0+0+ \; ... \; +0} \; wenn \; BF_c = BF_{s[j]} = 1 \end{cases}$$

Algorithmus - Step 3

$$\Sigma = W * V^{-sk}$$

V, W aus vorherigem Schritt einsetzen und für $pk = g^{sk}$

$$\Sigma = (g^{sk*s+r_{i_1}+r_{i_2}+ \; ... \; +r_{i_k}} * g^{-sk*s+r_{i_1}+r_{i_2}+ \; ... \; +r_{i_k}} * g^x)$$

$$\Sigma = g^x$$

The number of hash functions has distinct less influence

3 Results and Discussion

Both methods have shown to be capable to compare DNA datasets in an good time. Dauer für Vergleich des gesamten Exomes bei wenigen Minuten. Laufzeit Unabhängig davon wie stark die Überschneidung zwischen zwischen den Datensätzen ist. As shown in [**?**] the runtime is independent from the overlap of the two datasets

Überschneidung	14000	7500	5000	2000
Runtime (sec)	221	247	211	222
Abw. zur Überschn.	0.01%	3.3%	8.8%	36.8%

Table 1: Hashfunktionen : 14, Anzahl Bloomfilter Bits:3029660, Größe der Datensätze: 15000 SNPs

The runtime is lineary dependent to the amount of bloomfilter bits. The strenght of the deviation from the Die Laufzeit ist linear abhängig zur Anzahl der Bloomfilterbits Die Stärke der Abweichung ist ebenfalls linear abhängig zur Anzahl der Bloomfilter Bits

Array	1442696	1009887	577079	144270
Runtime (sec)	108	83	47	11
Abweichung	4%	6%	13%	51%

Table 2: Datensatz 1000 SNPs, Überschneidung 100, Hashfunktionen: 10

Hashf.	1	4	7	10	14
Runtime (sec)	7	27	44	62	104
Abweichung	11%	13%	10%	9%	9%

Table 3: Datensatz 1000 SNPs, Überschneidung 100, Array: 504944

Anzahl der Hashfunktionen hat deutlich weniger Einfluss, jedoch kommt es bei hoher Anzahl zu vermehrt Falsch positiven Ergebnissen.
The method based on paillier is significantly slower then the one based on elgamal. In fact the based on paillier needs smaller bloomfilters for the same accuracy, but the bitwise encryption takes way longer.
Zum Vergleich von gesamten Exomen ca. 40 min

3.1 Figures and Tables

Use the abbreviation "Fig." throughout the text of your manuscript, even at the beginning of a sentence. Do not abbreviate "Table." Tables are numbered with Arabic numerals, as can be seen in Table 7. Do not put borders around your figures. Each figure/table should be mentioned in the text.

Figure 1: It is good practice to explain the significance of the figure in the caption. Each figure should be able to be stand alone.

Figure axis labels are often a source of confusion. Use words rather than symbols. As an example, write the quantity "Energy," or "Energy, E," not just "E." Separate units with a slash, e.g. "Energy / J". Do not label axes only with units. To provide consistent reproducibility, please include axes and tick marks on all four sides of your graphs and avoid the use of grid lines (note that grid lines tend to clutter a graph if dark or reproduce poorly if light). Please also include an explanatory legend within your graphs when two or more curves or sets of data are included. Avoid explaining the different symbols and curves in the figure caption alone - using a legend results in a much more easily understood figure. Fig. 2 is a good example.

Array	14139	12119	10099	8080
Runtime (sec)	219	194	183	163
Abweichung	1%	4%	6%	24%

Table 4: Hashf.7, Überschneidung 100, SNPs 1000

Array	141385	100989	75742
Runtime (sec)	2420	2318	2007
Abweichung	1%	4%	13%

Table 5: Hashf.7, Überschneidung 7500, SNPs 15000

Please remember that the book of abstracts will be black and white. Make sure figures are still legible. Fig. 1 is a good example for a black and white figure.

3.2 References

Number citations consecutively in square brackets [1]. The sentence punctuation follows the brackets [2]. Multiple references [3], [4] are each numbered with separate brackets [2]-[5]. In sentences, refer simply to the reference number, as in [1]. Do not use "Ref. [2]" or "reference [3]" except at the beginning of a sentence: "Reference [4] shows ... ."
Please note that the references at the end of this document are in the preferred referencing style. Give all authors names; do not use "et al." unless there are six authors or more. Use a space after authors' initials [5]. Papers that have not been published, personal communication and comparable references are no full-value references. Try to avoid them.
Capitalize only the first word in a paper title, except for proper nouns and element symbols. For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [1].
Use only the bibliographystyle ieeetr. You can either include the bibliography in this document or you use an editor, e.g. JabRef.

3.2.1 Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have already been defined in the abstract. Abbreviations such as IEEE, SI, ac, and dc do not have to be defined. Abbreviations that incorporate periods should not have spaces: write "C.N.R.S.," not "C. N. R. S." Do not use abbreviations in the title unless they are unavoidable.

3.3 Equations

Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Use parentheses to avoid ambiguities in denominators. Punctuate equations when they are part of a sentence, as in

$$E = mc^2. \tag{1}$$

Abweichung	0.1%	0.6%	2%	3%	4%	6%
Runtime elgamal	467	150	17	15	11	6
Runtime paillier	510	340	150	150	135	120

Table 6: Hashf.7, Überschneidung 100, SNPs 1000

Table 7: Fontsize and Format

Element	Font Size	Font Type
Title	16 pt	Bold
Authors	10 pt	Normal
Abstract	10 pt	Normal
Heading 1	14 pt	Bold
Heading 2	12 pt	Bold
Heading 3	10 pt	Bold
Text	10 pt	Normal
Table Title	10 pt	Normal
Figure	10 pt	Normal
References	10 pt	Normal

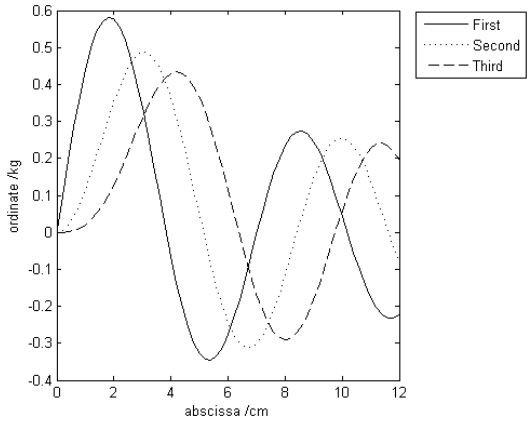

Figure 2: Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

Be sure that the symbols in your equation have been defined before the equation appears or immediately following. Refer to "(1)," not "Eq. (1)" or "equation (1)," except at the beginning of a sentence: "Equation (1) is ... ".

4 Conclusion

Please note: If your university supervisor is not named as an author, you should change the acknowledgement to: The work has been carried out at Genius Industries, Valley of Innovation and supervised by D. Musterprofessor, Institute of Wisdom, Universität zu Lübeck.

Acknowledgement

The work has been carried out at Genius Industries, Valley of Innovation and supervised by the Institute of Wisdom, Universität zu Lübeck.

As an option, additional acknowledgements can be added.

5 References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*. Addison-Wesley, Harlow, 1999.

[2] BMT 2018 Aachen, *Save the Date*. Available: https://www.vde.com/en/events/bmt [last accessed on 2017-11-30].

[3] M. Young, *The Technical Writer's Handbook*. University Science, Mill Valley, 1989.

[4] C. Kaethner, J. Müller and T. M. Buzug, *Phantom-based Determination of Noise Distribution in Computed Tomography*. In: Student Conference on Medical Engineering Science 2012, Grin Publishing, München, pp. 59–62, 2012.

[5] D. Zongker, *Chicken Chicken Chicken*. Annals of Improbable Research, vol. 12, no. 5, pp. 16–21, 2006.