A Metaphone based Chaotic Searchable Encryption Algorithm for Border Management

Abir Awad[1,2] and Brian Lee[1]

[1]*Irish Centre for Cloud Computing and Commerce (IC4), Athlone Institut of Technology, Athlone, Ireland*

[2]*Facuty of Computing, Engineering and Science, University of South Wales, Pontypridd, U.K.*

Keywords: Privacy, Personal Data, Searchable Encryption, Fuzzy Search, Chaotic Locality Sensitive Hashing, Metaphone.

Abstract: In this paper, we consider a use case for national border control and management involving the assurance of privacy and protection of personally identifiable information (PII) in a shared multi-tenant environment, i.e. the cloud. A fuzzy searchable encryption scheme is applied on a watch list of names which are used as indexes for the identification files that are in their turn encrypted and stored on the cloud. Two propositions are described and tested in this paper. The first entails the application of a chaotic fuzzy searchable encryption scheme directly on the use case and its subsequent verification on a number of phonetics synonyms for each name. In the second version, a metaphone based chaotic fuzzy transformation method is used to perform a secure search and query. In this latter case, the fuzzy transformation is performed in two stages: the first stage is the application of the metaphone algorithm which maps all the words pronounced in the same way to a single code and the second stage is the application of the chaotic Local Sensitive Hashing (LSH) to the code words. In both the first and second propositions, amplification of the LSH is also performed which permits controlled fuzziness and ranking of the results. Extensive tests are performed and experimental results show that the proposed scheme can be used for secure searchable identification files and a privacy preserving scheme on the cloud.

1 INTRODUCTION

Privacy–respecting fuzzy matching is a key requirement in many applications e.g. public health, biomedical research, border control management etc. (Hoepman, 2006) - (Bissessar et al., 2016). Matching words and names that sound similar is important for border security and for many other applications. However, outsourced personal information and identity files to the cloud or any shared domain need also to be secured/encrypted.

A border management text based "watchlist" search would also need to cater for Phonetic String Matching which is a kind of non-exact (fuzzy) text searching requirements.

This should cover the following type of errors:

- Phonetic equivalent spelling variants: e.g. Alain, Alan, Allain, Allan, Allen, Allin, Allyn.
- Transliteration spelling differences (e.g. from Arabic to Latin script): e.g. Muammar Gadafi, Moamar Gaddafi, Mo'ammar Gadhafi, Muammar Gathafi, Muammar Ghadafi.

In this paper, we propose to use the adapted version of the chaotic fuzzy searchable encryption proposed in (Awad et al., 2015) for this use case. Searchable encryption is a scheme that allows to search over encrypted files by the device of keywords. In this approach the keyword, or index, is the name of the person that will indicate his full identity file which is stored on the cloud in an encrypted format. Recently, many approaches have been proposed to enable fuzzy search. Wildcards were used for similar keywords search (Li et al., 2010). However, this technique only covers part of the possibly nearby keywords (Yang et al., 2014). Other fuzzy search approaches using the locality sensitive hashing (LSH) were also proposed recently for secure storage and search on the cloud (Bringer and Chabanne, 2011), (Kuzu et al., 2012), (Awad et al., 2014). In (Awad et al., 2015), we proposed a chaotic searchable encryption for a secure cloud storage. This method was proposed for a mobile cloud storage use case scenario for secure storage and retrieval of content based text files. In this paper, we

Awad, A. and Lee, B.
A Metaphone based Chaotic Searchable Encryption Algorithm for Border Management.
DOI: 10.5220/0005953503970402
In *Proceedings of the 13th International Joint Conference on e-Business and Telecommunications (ICETE 2016) - Volume 4: SECRYPT*, pages 397-402
ISBN: 978-989-758-196-0

apply two modified versions of this algorithm for the border management use case. In our new scenario, there is no need for a posting list or for order preserving encryption. Each keyword is, in this case, the name of each person and is used as a reference to one identity file. When querying with a slightly erroneous name, the algorithm retrieves the best matched names and ranks them using the similarity of the stored fuzzy transformation values and return these results, names and identity files, to the requester (border management) who can perform any further investigation, if needed, to find out if the person crossing the border is on the watch list or not.

This paper is organised as follows: Sections 2 gives some background information. Section 3 and 4 describes respectively the proposed and tested chaos based searchable encryption algorithms with and without the metaphone matching transformation. Section 5 presents the test bed and the simulation results. Finally, Section 6 summarizes our conclusions.

2 BACKGROUND

2.1 Locality Sensitive Hashing

LSH functions reduce with high probability the differences occurring between similar data i.e. similar results are obtained for data with close proximity but distant data remain remote.

Let B be a metric space, $r_1, r_2 \in \mathbb{R}$ with $r_1 < r_2$ and $p_1, p_2 \in [0,1]$ with $p_1 > p_2$

A family $H = \{h_1, \dots, h_\mu\}$ is an LSH family if for all $x, x' \in B$:

$$\Pr_H[h(x) = h(x')] > p_1, \text{if } d(x, x') < r_1 \quad (1)$$

$$\Pr_H[h(x) = h(x')] < p_2, \text{if } d(x, x') > r_2 \quad (2)$$

d is the distance metric utilized (e.g. Hamming distance).

Jaccard coefficient is usually used to measure the similarity between two sets A and B containing words from two documents (Awad et al., 2015). It is defined in (3) as follows:

$$s(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

The distance between these sets can be obtained by (4) as follows:

$$d(A, B) = 1 - s(A, B) \quad (4)$$

Min-hash or Min wise independent permutations is one of the most used LSH functions. If π is a random permutation, the hash value is defined by (5) as follows:

$$h_\pi(A) = \min\{\pi(a) \,/\, a \in A\} \quad (5)$$

and the probability that the two hashed values are equal, is equal to the Jaccard distance (6):

$$\Pr[h_\pi(A) = h_\pi(B)] = s(A, B) \quad (6)$$

In our proposal, min-hash is used to support the fuzzy transformation applied on the names indexing the outsourced files.

2.2 Amplified Minhash Methods

To amplify a locality-sensitive hashing family a AND-OR construction can be used (Kuzu et al., 2012).

The AND construction is formed with k random functions from H: $g_i = \left(h_{i_1} \Lambda h_{i_2}, \dots \Lambda\, h_{i_k}\right)$. In this context, $g_i(x) = g_i(y)$ if and only if $\forall j \left(h_{i_j}(x) = h_{i_j}(y)\right)$ where $1 \le j \le k$. The OR construction is formed with λ different AND constructions such that $g(x) = g(y)$ if and only if $\exists i\left(g_i(x) = g_i(y)\right)$ where $1 \le i \le \lambda$. With such a construction, we can turn an (r_1, r_2, p_1, p_2) sensitive family into an (r_1, r_2, p'_1, p'_2) sensitive family where $p'_1 = 1 - \left(1 - p_1{}^k\right)^\lambda$ and $p'_2 = 1 - \left(1 - p_2{}^k\right)^\lambda$.

Amplified LSH is used in our proposal to improve the results of the fuzzy transformation step of our searchable encryption algorithms.

2.3 Chaotic Minhash Methods

Chaos has a number of interesting properties, e.g. good pseudo-randomness and sensitivity to its control parameters that can be directly linked to the properties of confusion and diffusion in cryptography. In addition, these systems are deterministic, meaning that their future behavior is fully determined by their parameters, with no random elements involved. However, the chaotic signal is pseudo-random and may appear as noise for unauthorized users. The idea of taking advantage of digital chaotic systems and of constructing chaotic cryptosystems has been extensively investigated and attracted many researchers (Awad, 2010) – (Rostom et al., 2014) but very few researchers considered using the chaos for searchable encryption. In (Awad et al., 2015), we proposed new minhash methods based on Piece Wise Linear Chaotic Map (PWLCM).

In this method, the translation i.e. the encoding of the keyword, is performed by the chaotic map instead of the Bloom filter used by Kuzu et al. (Kuzu et al., 2012). PWLCM is then used to transform the keyword to a set of numbers that will be used as input for the minwise permutation method in order to obtain finally the minhash value.

A 1-gram shingling is applied on each name and the ASCII code of each letter is mapped to the interval [0,1] and then encoded by the chaotic map. For each shingle, a number of iterations are performed and the obtained chaotic values are then mapped to integers in the interval [0,*m*], where *m* is a secret parameter for the minhash. Finally, the keyword is represented by an array of values that are used as an input for the minhash method. The amplified chaotic minhashes are finally obtained by applying the amplification method i.e. the AND-OR construction on each one in addition with chaos.

The amplified chaotic minhash is used in the fuzzy transformation stage of our proposed searchable encryption algorithms.

2.4 Metaphone

A phonetic algorithm is an algorithm to identify words with similar pronunciation and is used to index the words based on their pronunciation. The first metaphone or a phonetic algorithm was published by Lawrence Philips (Lawrence, 1990) in 1990 and it was used for indexing words by their English pronunciation. A new version of the algorithm, which is named Double Metaphone (Lawrence, 2000) is later produced and it did take into account spelling peculiarities of a number of other languages in addition to the English e.g. French, Italian, Spanish, Chinese… In 2009, Lawrence Philips released a third version, called Metaphone 3, which achieves an accuracy of approximately 99% for English words, non-English words familiar to Americans, and first names and family names commonly found in the United States. All these algorithms were basically built to discover similar pronunced names stored in large databases (Parmar and Kumbharana, 2014). These algorithms are slightly different but a metaphone algorithm usually operates by first removing non-English letters and characters from the word being processed. Next, all vowels are also discarded unless the word begins with an initial vowel in which case all vowels except the initial one are discarded. Finally all consonants and groups of consonants are mapped to their Metaphone code.

The phonetic algorithm is used to improve the fuzziness in our proposed metaphone based searchable encryption algorithm.

3 SEARCHABLE ENCRYPTION ALGORITHM

The proposed approach allows to search over encrypted identity files stored in the cloud and returns the relevant files to the queries in a ranked order. This scheme permits search not only with the exact index name used during the cloud storage process, but also with an approximate keyword i.e. a misspelling name.

It consists of two different phases: the storage phase and the search phase. In the storage phase, the border management creates, from the watch list names, the meta-data necessary for the cloud provider to search the full identity files of these people. Then, they encrypt these files and store them in the shared environment i.e. the cloud. In the search phase, when a person is trying to cross the border, the responsible person on the border control queries the cloud to find out whether this person is on the watch list or not. The cloud receives the hashed query, performs the search, retrieves and returns the matched identity files in a ranked order based on their similarity to the query. We give below the detailed description for the both processes.

3.1 Storage Phase

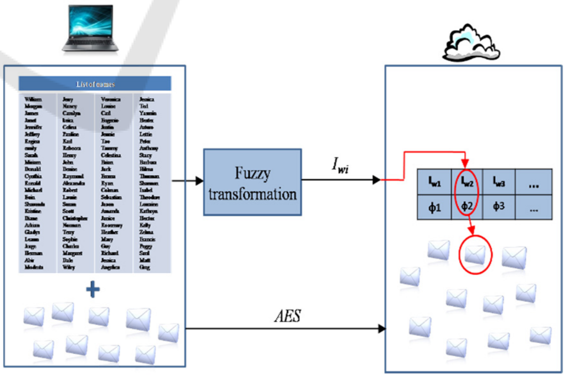

Figure 1: Storage of the names and identities on the cloud.

This phase consists, basically, of two parts (see Fig. 1); the fuzzy transformation of the names i.e. keywords and the encryption of the actual identity files. The obtained index and encrypted files are then stored in the cloud. We used AES for the encryption of the identity files and the amplified chaotic local sensitive hashing explained in section 2 for the fuzzy transformation.

3.2 Search Phase

During this process, the responsible on the border control needs to perform the fuzzy transformation on the name of the suspected person and query the cloud with it. In its turn, the cloud uses this index (hashed name) to retrieve the most relevant files for this name/query (see Fig. 2).

We assume that the user is querying with a slightly different name than the stored one. In our test, we consider that the user is querying by one of the phonetic synonyms of the original name or by a misspelling name. Then the secure search scheme is applied on this erroneous name and the query is sent to the cloud. This later searches over the stored data and return the identities of the most similar names to the query which will be then decrypted on his side. In our algorithm, the user can specify the maximum number of identity files that he wants to receive from the cloud.

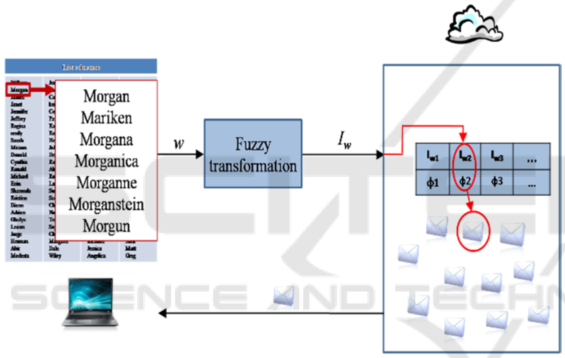

Figure 2: Search with the phonetic synonyms on the cloud to find the identity files.

4 METAPHONE BASED SEARCHABLE ENCRYPTION

In this section, we explain the second proposed searchable encryption algorithm for the border management. This algorithm is using an improved fuzzy transformation by applying the amplified chaotic LSH on the code obtained by the metaphone algorithm and not directly on the actual name as performed in the algorithm explained earlier in section 3.

4.1 Storage Phase

The double fuzzy transformation is applied on each name and then the index is then sent to the cloud along with the encrypted (using AES) identity file. The cloud provider then stores this information into a hashmap using the secure index (encrypted name) as a reference for the corresponding encrypted identity files. The fuzzy transformation in this case is a combination of the metaphone algorithm and the amplified local sensitive hashing (see Fig. 3).

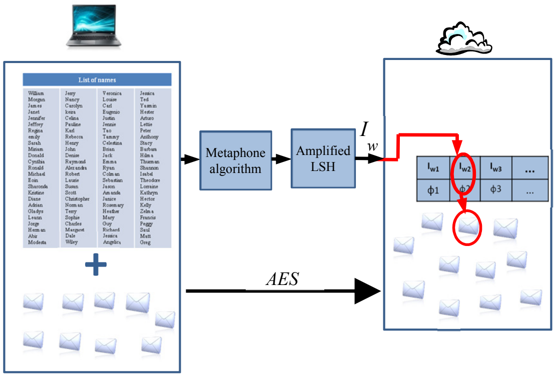

Figure 3: Storage of the names and identities on the cloud for the metaphone based searchable encryption.

4.2 Search Phase

Similar to the search phase of the previous searchable encryption algorithm we assume that the border control person is querying by the exact name or by one of the phonetic synonyms of the original name or by a mispelled name. The same metaphone algorithm followed by the amplified chaotic LSH are applied on this keyword. The cloud then uses this index to retrieve the most relevant ID files (see Fig. 4).

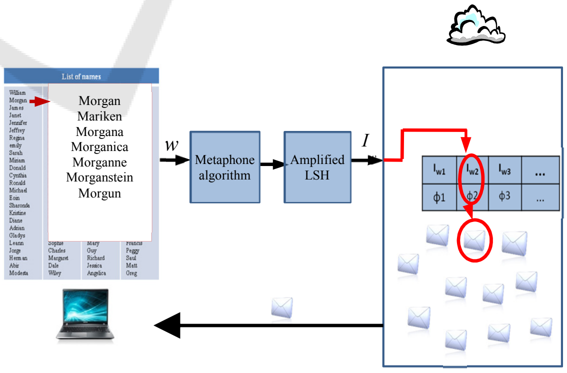

Figure 4: Search with the phonetic synonyms on the cloud to find the identity file.

5 EXPERIMENTAL RESULTS

All of our programs are written in Java. To apply the algorithm on the identities use-case, a test-bed needed to be prepared. A list of 100 distinct (fake) names are used in our test and a metaphone file which contains the phonetic synonyms for each name is created.

Finally, the two proposed versions of the fuzzy searchable encryption are applied on these synonyms to prove the efficiency of the proposed methods.

5.1 Testbed Preparation

5.1.1 Generation of a List of Fake Names

We generated a list of names using a fake name generator (fakenamegenerator, 2016). This generator is first used to generate a file containing 100 fake identities then these names and the corresponding identities are inserted into different files for testing purpose. We avoided the repetition of names in the created list which means that each name appears once in the list to be tested.

5.1.2 Generation of Phonetic Synonyms

In this section, we explain how to generate the phonetic synonyms for each of the names for testing purpose i.e. to query with erroneous/similar names. To generate the phonetics synonyms for each name, we did follow the following procedure which consisted on two stages: pre-computation and phonetics hashmap creation.

Pre-computation

The goal of this phase is to create a "metaphone hashmap" where a user can query with a metaphone code and find all the words which might have similar pronunciation.

Fig. 5 shows how this hashmap is created.

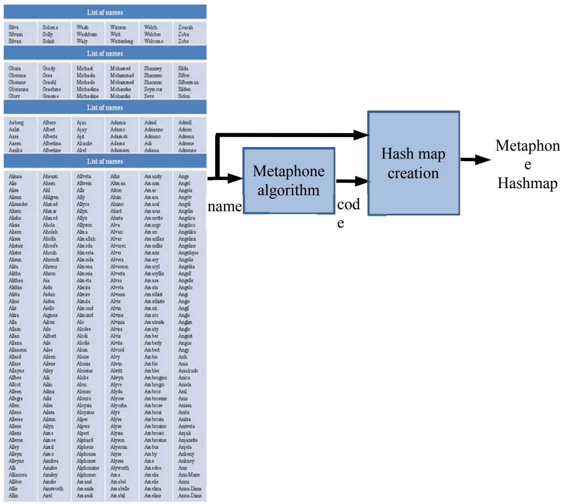

Figure 5: Metaphone hashmap creation.

A word list of names is taken from the "Moby Words" project (Moby Words, 2016) which is a collection of public-domain lexical resources created by Grady Ward in 1996. The used list contains the most common 21986 names in the United States and Great Britain. The metaphone algorithm is then applied on these names and a hashmap is then created. Each metaphone code is then a reference to a number of names which we called the phonetics synonyms.

Test Phonetics Synonyms Generation Phase

As shown in Fig. 6, to find the list of words that might have a similar pronunciation for each of our test names, we applied the metaphone algorithm, found the metaphone code and queried with it on the metaphone hashmap (computed in the pre-computation phase) to find the phonetic synonyms for each test name. Then, a metaphone file containing the phonetics synonyms is created for each name and is then used for the testing of our algorithms. This size of this file varied from a name to another depending on the number of phonetic synonyms found in the metaphone hashmap.

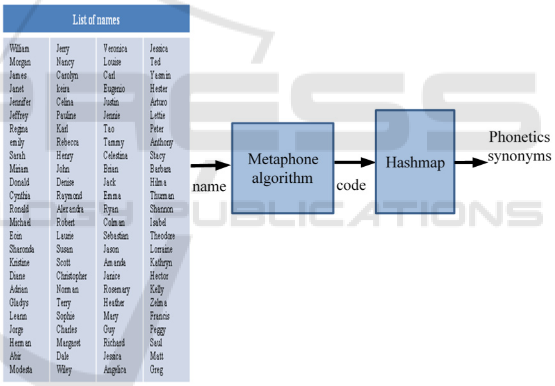

Figure 6: Phonetics synonyms generation.

5.2 Test Results

In order to test our proposed approaches, we queried with each phonetic synonym, generated earlier for testing purpose, and we calculated the percentage of success of retrieving the right identity file for each name. Then, we calculated the average for the 100 names.

In the first test, we assumed that the user is receiving just one identity file from the cloud. The obtained average of success over the 100 names is equal to 0.793 for the first version of the searchable encryption method and 0.99 for the metaphone based searchable encryption. As we can see, the metaphone based searchable encryption method is more successful in retrieving the required identity file, when querying with a misspelling name, comparing to the original chaotic searchable encryption method.

The same test is then performed when the user did require the most similar three identity files for his query and the results for the first and second algorithms are then calculated. The averages over the 100 names for the first and second algorithms are respectively equal to 0.796 and 1. As we can see, the algorithms are more successful, in this case, to find the right identity files when querying with a misspelling name. The main reason is that the cloud provider is retrieving three files and not just one and then there is a bigger chance to find the matched identity file.

6 CONCLUSIONS

In this paper, we proposed the first combined chaos and metaphone based searchable encryption approach. The proposed algorithm allows fuzzy keyword searches over the encrypted data stored on the cloud. Our approach proved the possibility of the usage of the searchable encryption on the identity storage use case and guarantees the privacy and confidentiality of the people crossing borders even *vis*-à-*vis* the cloud provider who is semi-trusted in our case.

ACKNOWLEDGEMENTS

This work is made within the framework of the Irish Centre for Cloud Computing and Commerce (IC4) which is an Irish government and Enterprise Ireland supported technology research centre established in 2012.

REFERENCES

Awad, A., Matthews, A., Qiao, Y., Lee, B., 2015. "Chaotic Searchable Encryption for Mobile Cloud Storage." *IEEE Transactions on Cloud Computing.*

Li, J., Wang, Q., Wang, C., Cao, N., Ren, K., Lou, W., 2010. "Fuzzy keyword search over encrypted data in cloud computing," *INFOCOM, 2010 Proceedings IEEE*, Dept. of ECE, Illinois Inst. of Technol., Chicago, IL, USA.

Yang, B., Pang, X., Du, Q., Xie, D., 2014. "Effective Error-Tolerant Keyword Search for Secure Cloud Computing," *Journal of computer science and technology*, vol. 29, no.1, pp. 81-89.

Bringer, J., Chabanne, H., 2011. "Embedding edit distance to enable private keyword search," Secure and Trust Computing, Data Management and Applications, *Communications in Computer and Information Science*, vol. 186, no. 1, pp. 105-113.

Kuzu, M., Islm, M. S., Kantarcioglu, M., 2012. "Efficient similarity search over encrypted data," ICDE's12 proceedings of the 2012 *IEEE 28th International conference on data engineering*, pp. 1156-1167, IEEE computer society Washington, DC, USA.

Awad, A., Awad, D., 2010. "Efficient image chaotic encryption algorithm with no propagation error," *ETRI journal*, vol. 32, no. 5, pp. 774-783.

Awad, A., Miri, A., 2012. "A new image encryption algorithm based on a chaotic DNA substitution method," in Communications (ICC), 2012 IEEE International Conference on, pp. 1011-1015.

Ismail, M., Chalhoub, G., Bakhache, B., 2012. "Evaluation of a fast symmetric cryptographic algorithm based on the chaos theory for wireless sensor networks," in *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*, pp. 913-919.

Rostom, R., Bakhache, B., Salami, H., Awad, A., 2014. "Quantum cryptography and chaos for the transmission of security keys in 802.11 networks," in *the 17th IEEE Mediterranean Electrotechnical Conference (MELECON)*, pp. 350-356.

Lawrence, Ph., 1990. "Hanging on the Metaphone," *Computer Language*, Vol. 7, No. 12.

Lawrence, Ph., 2000. "The double metaphone search algorithm." *C/C++ users journal,* pp. 38-43.

Metaphone3. http://www.amorphics.com/ metaphone3.html.

Parmar, V. P., Kumbharana, C. K., 2014. "Study Existing Various Phonetic Algorithms and Designing and Development of a working model for the New Developed Algorithm and Comparison by implementing it with Existing Algorithm". *International Journal of Computer Applications*.

Fakenamegenerator, 2013. http://www.fakenamegenerator.com/ order.php.

Moby Words, 2016. "Moby Words" project. http://icon.shef.ac.uk/Moby/mwords.html.

Awad, A., Matthews, A., Lee, B., 2014. "Secure cloud storage and search scheme for mobile devices," in *the 17th IEEE Mediterranean Electrotechnical Conference (MELECON)*, pp. 144-150.

Hoepman, J. H., Hubbers, E., Jacobs, B., Oostdijk, M., Schreur, R W., 2006. "Crossing borders: Security and privacy issues of the european e-passport," In *Advances in Information and Computer Security, Springer Berlin Heidelberg*, pp. 152-167.

Pang, C., Gu, L., Hansen, D., Maeder, A., 2009. "Privacy-preserving fuzzy matching using a public reference table," In *Intelligent Patient Management, Springer Berlin Heidelberg*, pp. 71-89.

Bissessar, D., Adams, C., Stoianov, A., 2016. "Privacy, Security and Convenience: Biometric Encryption for Smartphone-Based Electronic Travel Documents," In *Recent Advances in Computational Intelligence in Defense and Security 2016, Springer International Publishing,* pp. 339-366.