Applying quantum machine learning approach for detecting chaotically generated fake usernames of accounts

Desislav Andreev, Simona Petrakieva, Ina Taralova, Zongchao Qiao

Applying quantum machine learning approach for detecting chaotically generated fake usernames of accounts

Security systems with heightened safety of information

Desislav Andreev
Dep. Computer Systems,
Faculty of Computer Systems and Technologies
Technical University of Sofia
Sofia, Bulgaria
desislav.andreev@gmail.co

Ina Taralova
Laboratoire des Sciences du Numérique de Nantes
LS2N, Ecole Centrale de Nantes
Nantes, France
ina.taralova@ec-nantes.fr

Simona Petrakieva
Dep. Theory of Electrical Engineering,
Faculty of Automation
Technical University of Sofia
Sofia, Bulgaria
petrakievas-te@tu-sofia.bg

Zongchao Qiao
Laboratoire des Sciences du Numérique de Nantes
LS2N, Ecole Centrale de Nantes
Nantes, France
qiaozongchao@163.com

Abstract—**The present paper is further development of our previous publication about revealing false usernames of accounts as a result of hackers' attacks. Although both papers use methods for machine learning analysis, the novelty in this publication consists in applying the quantum technique for making cluster analysis and a new chaotic generator producing fake usernames with unfixed length. The consequence of actions necessary for analyzing the names of users' accounts is following. First, a chaotic pseudo random number generator (PRNG) producing aperiodic time series is proposed. Following given rules, the latter are used to produce fake accounts usernames used as a data base to test the efficiency in avoiding malicious intentions. The generated names are with low, middle and high randomization. Second, these names feed as an input to the quantum machine learning algorithm, which divides them in different clusters. Next, the Quantum Silhouette algorithm for quality evaluation of the results from clusterization, is applied. In the end of the paper, the suggested new technique – chaotic generator and quantum clustering algorithm – is illustrated on the illustrative example including 100 000 usernames of accounts.**

Keywords-chaotic pseudo random number generator (PRNG), quantum machine learning clustering, k-means, Silhouette

I. INTRODUCTION

Nowadays the information is the most esteemed and the most expensive thing in the world. That's why it has a key role in the relationships between the humans. For that reason the information has to be protected from unwanted hackers' attacks. Internet is enough aggressive environment and web systems are always attacked by malicious users. Users of the arbitrary electronic system have own private account, which is secured by username and password. But do they always take the necessary preventive measures for certain safeguard of their information against unwanted hackers' attacks? For this reason we apply some other additional methods for protection the users' accounts from illegal hackers' access. There exist a lot of methods for this protection, which are developed and sold as a software products in the market. For example:

ConsentIQ [1] is a solution which provides an easily implemented privacy regulation for any small business. It captures private data from EU citizens, which specifies its work with EU General Data Protection Regulation (GDPR) and the pending ePrivacy regulation \-sometimes called the 'cookie' law).

Dhound [2] is orienting on the protection against legitimate users with malicious intentions. It tracks and alerts successfully suspicious events, controls output traffic that allows to notify data owners quickly that somebody without any reason accessed the server or your server suddenly started working with unknown services in Internet. Dhound IDS does not require specific security knowledge to manage.

Other example about the efforts of the people for secure their information are measures made from the German Association for Data Protection [3]. The latter advising medium-sized companies and corporations in all aspects of data security and protection of their personal data.

In the present paper we propose a new technique for detecting the false usernames of accounts. It combines the properties of the chaotic generator for generating the usernames and their false derivatives, which next divide in

separately clusters using quantum machine learning clustering method.

Chaotic generators have been attracting increasing attention due to their excellent random properties: each different starting point gives rise to different output sequence. In the same time, they exhibit typical deterministic properties, such as respectively: identical initial conditions will give rise to identical output sequences [4, 5].

Generally speaking, these generators can be successfully applied as Pseudo-Random Numbers Generators, and have excellent features, as for randomness, spectrum etc. [6, 7]. However, many open problems arise, such as the required structure for an efficient chaotic generator (e.g. the chaotic map), the criteria to analyze the tuning parameters, the choice of the best coupling in order to satisfy the predefined criteria for strong chaoticity, etc.

Quantum machine learning technique is a unification between Quantum Physics (QP) and Machine Learning (ML) algorithms. It uses the ideas of the classical ML algorithms to analyze quantum systems [8]. After applying Quantum machine learning technique the resulted clusters are examined through the Silhouette method [11]. So in briefly, in this paper, we do the following:

1) Generating the database of complex different names by Pseudo Random Chaotic Generator;

2) Clusterization by a quantum machine learning Technique;

3) Quality assessment of the resulting clusters by Silhouette method.

The present paper is organized as follows. In the next section II the problem to reveal false users' account is defined. It also describes the chaotic generator producing false names fed to the quantum machine learning algorithm. The main requirements for chaotic generator design are proposed in section III. In section IV the functionality of the new suggested quantum clustering technique for separating the names from chaotic generator, is described in details. This technique is applied in section V on the illustrative example including 100 000 names. The clustering results also are compared there with those obtained from the Silhouette method. The paper finishes with conclusion remarks about advantages and disadvantages of the new quantum machine learning clustering algorithm working with the false usernames, generated from the proposed pseudo random chaotic generator.

II. Problem Statement

Generated usernames will have the following features: *n* number of symbols (n > 6) where *n* is a combination of letters and alphanumerical symbols. UTF-8 coding has been used. The number of usernames is 100 as each of them will be in 1000 variations, i.e. totally 100 100 usernames in total will be clustered by a quantum machine learning algorithm.

In this particular application, some specific characteristics of the chaotic generator are required, so that it can be successfully used as a pseudo-random number generator with independent output sequences. Note that not all chaotic maps are suitable for this implementation. In particular, the chaotic generator is considered random, if it passes successfully the NIST tests for randomness, and if the generated output values are equidistributed, so they should all appear with the same probability. This is very important, because our laws for fake usernames generation are based on the assumption that all rules from 0 to 9 (taken as the first digit of the generated pseudo-random value) have the same chance to be drawn.

Quantum machine learning technique is expected to be faster for clustering than the related classical representation of k-means, Mini-batch k-means, agglomerative clustering, etc. [9, 10].

III. Methodology for Synthesis of the Random Chaotic Generator

In this paper we have used random law to generate the fake usernames from the real user data base containing a pre-specified number of real usernames with not fixed length (6 symbols minimum). Note that in this particular application, it is not necessary to binarize the chaotic generator output as we deal with real values. Binarization is often a tricky point, since there is not a strict rule, and quantization can lead to loss of randomness etc.

The algorithm that we have defined for fake usernames generation is the following. According to the random generated value, different (specific) rules have been applied to alternate the symbols of each real username: remove, double, swap, add, shift, and combinations of the above. Additional test have selected symbols with higher frequency in the data base (e.g. vowels, etc.). As a result, 1000 fake usernames have been created for each real username, and served as input for the cluster analysis.

IV. Basis of the Quantum Machine Learning Technique

In The quantum machine learning technique is a new approach for clustering. It combines the Quantum Physics (QP) and Machine Learning (ML) in one algorithm. The originality of the new methodology consists in the fact that it uses classical ML algorithms applied in QP.

Before we start with the main concepts of the quantum representation transformation actions, we need to summarize the three concepts existing in the Quantum Mechanics that could be counted as prerequisites to the current research. In the first place this is the quantum parallelism [9], which unites the *superposition* principle [9] and the linearity of the quantum world in order to evaluate a single function simultaneously on arbitrarily many inputs. Immediately follows the quantum *interference* [9], that makes possible the logical paths from a given execution to affect each other in a positive way. This is actually a welcomed, so-to-say, output, because the influence between the positive outputs could affirm one another and the negative would fall out at certain point. This self-rearrangement of the system is typical expectation of the machine learning, especially when it comes to evaluations of algorithms over big data [9]. But there are also quantum states

that are multi-particle, which cannot be described by the independent state of a single particle. Correlation between these states cannot be examined the classical way and therefore it is required to postulate the main source of the quantum information analysis and a powerful communication resource – the quantum *entanglement* [9].

The algorithms should be able to execute correctly the tasks for *pattern classification, pattern recognition* - to assign labels correctly to a given input set or finding a shape of patterns; *pattern completion* - adding missing values in the input dataset; *associative memory* - retrieving one of a number of stored memory vectors upon an input. Let's focus currently on the first one: If we take a look at the classification and clustering problems in the classical machine learning, it will be clear that the main concern is in finding the efficient calculations of the classical distances on a potential quantum computer. This is obviously required for the similarity measurement of two feature vectors. For the classical representation the Euclidian or the Hamming distances are used. For the quantum world there are few suggestions from [9]. All of them are basing their similarity measure on the overlap or fidelity $|\langle a|b\rangle|$ of two quantum states |a□ and |b□, respectively. The *swap* test is proposed, where |a, b, 0_{anc} □state with the two wave-functions and one ancilla register - described in [9], set to 0 is initially provided and fed to a Hadamard gate (Figure 1):

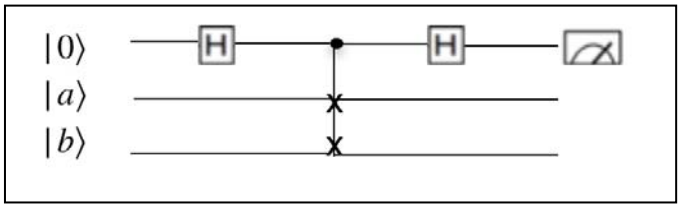

Figure 1. *Swap* test *logical circuit presentation*

The following Hadamard transformation sets the ancilla into a superposition $\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. The SWAP-gate on *a* and *b* swaps the two states if the ancilla is in $|1\rangle$. The second Hadamard gate results in state:

$$|\psi_{sw}\rangle = \frac{1}{2}|0\rangle(|a,b\rangle + |b,a\rangle) + \frac{1}{2}|1\rangle(|a,b\rangle - |b,a\rangle) \quad (1)$$

for which the probability of measuring the ground state is given by:

$$P(|0_{anc}\rangle) = \frac{1}{2} + \frac{1}{2}|\langle a|b\rangle|^2 \quad (2)$$

the probability of 1/2 means that the two states do not overlap (they are orthogonal), while probability 1 indicates that they have maximum overlap.

V. Applying Quantum Machine Learning Technique for Reveal the False Names of Users' Accounts Generated by Chaotic Generator

Proceeding further we focus our experiment on the clustering algorithms - *k-means* to be precise, because it uses that distance measure we discussed above. It is also easy to be represented mathematically both in classical and quantum machine learning. The method is NP-hard, which means it is computationally difficult, which gives further stimulation to the quantum approach - this will be discussed later. Its overall classic complexity is *O(kNI)*, where *k* is the number of clusters, *N* is the number of samples and *I* – the number of iterations. The algorithm is alternating constantly between two main steps - assign step and update step:

- **Assign** each observation x_p to the cluster whose mean has the least squared Euclidean distance (nearest mean) and the observation is assigned to exactly one $S(t)$:

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \le \|x_p - m_j^{(t)}\|^2 \ \forall j, 1 \le j \le k\} \quad (3)$$

- **Update** (minimize the average distance) the means to be the centroids of the observations in the new clusters:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|}\sum_{x_j \in S_i^{(t)}} x_j \quad (4)$$

It stops when the assignments no longer change. The main idea is to partition *N* observations into *k* clusters. In the end the data space is partitioned in Voronoi cells. If the measurement space is *D* dimensional the time complexity of the algorithm would be *O(D)*, where each step takes time $O(N^2D)$. Having the observations in section IV above it would be expected, that the algorithm takes $O(Nlog(ND))$. Both times *N* is required at least once, because every vector is tested individually for the reassignment at each step [9]. Let's construct the state:

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|\vec{x}\rangle|0\rangle + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}|\vec{y_i}\rangle|i\rangle), \quad (5)$$

where *S* is the current cluster for a set of *N* reference vectors $\{|\vec{y_i}\rangle\}$ of length *P* and input vector $|\vec{x}\rangle$. The formulation of this equation results from (3). The distance can be efficiently calculated within error $\epsilon = O(\epsilon^{-1}\log NP)$. Getting back to the *swap* test and apply it over the previous step, we construct the following:

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|x\rangle|0\rangle - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}|y_i\rangle|i\rangle), \quad (6)$$

where

$$Z = |x|^2 + (\frac{1}{N})\sum_j |y_j|^2$$

This is repeated for each cluster until a desired confidence is reached, which is usually noted by the person, working with

the system. This situation could be avoided by a usage of an algorithm that evaluates this confidence and marks the final iteration of the quantum *k-means.* Such method is the *Silhouette* method, which we already used in [9] and [12], and proved, that it the *k-means* assigns well the received data and has a high confidence. The formula is the following:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}, \qquad (7)$$

where *a* is the average distance of *i* to the points in its cluster and *b* - the minimal average distance of *i* to points in another cluster, e.g. *b* is the average dissimilarity between point *i* and the points in the closest cluster to its cluster. From the equation in Section IV and [9] we conclude that the average distance is actually the average fidelity between the points in the cluster, in our case the average fidelity between the state of the *i*-th variable in the *n*-sized cluster, which brings us to:

$$a(i) = \frac{1}{n_c} \sum_{i \neq j, j=1}^{n_c} \langle F(\rho_i, \rho_j') \rangle_{\rho_0} \qquad (8.1)$$

$$a(i) = \frac{1}{n_c} \sum_{i \neq j, j=1}^{n_c} \langle F(\rho_i, \rho_j') \rangle_{\rho_0} \qquad (8.2)$$

Remembering that: the fidelity of two state is $\langle F \rangle = \langle F(\rho, \rho') \rangle_{\rho_0}$, where ρ and ρ′ are the density matrices of the *i*-th variable and any other one, respectively over all initial states ρ_0 [9]. Further: *s* is the number of clusters, while *c* is the current one, n_S is the number of elements in the cluster *S,* where *i* doesn't belong to them, e.g. in S_c for the context of *b*. From the calculation of the *Silhouette* coefficient for all elements of $\vec{x}$ would be:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, for \forall i \subset \vec{x} \qquad (9)$$

The workability of the suggested in the previous two sections technique for reveal false users' names is illustrated on the example includes 100 names in 1000 variations, generated by the chaotic generator (see section III). Next, these names are clustered from quantum machine learning algorithm (see Section IV), which results' quality (confidence) is evaluated with the Quantum Silhouette method [9].

The dataset consists of 140 000 samples. Our initial experiment goes through the whole process as discussed above in order to distinguish the number of clusters, where the confidence of the clustering algorithm is at its highest point. On Figure 2 (here below) we can observe that the confidence drops drastically with higher values of *k* – placed on the *x-axis;* the results from the Silhouette method are placed on the *y-axis.*

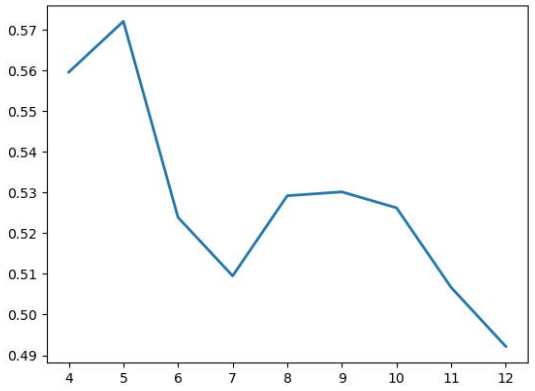

Figure 2. Results from the Silhouette evaluation for different number of clusters

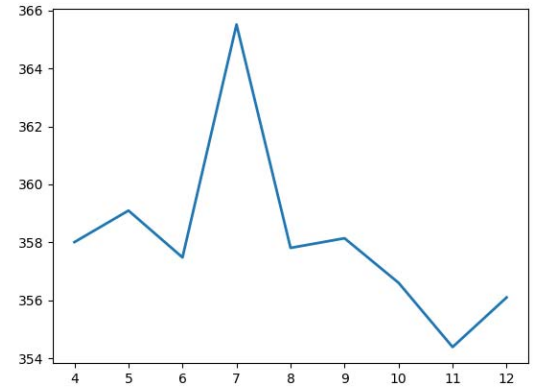

Figure 3. Time of execution for each value of k

Furthermore, we have to take into account the time of the execution for different number of clusters. Now we take a part of the dataset, around 10 % of it, in order to see the trend of the evaluation times for different number of clusters. On Figure 3 above the time values, measured in *seconds,* are placed on the *y-axis* of the graph and again – the value of *k*-s is on the *x-axis.* The outlying values for $k = 7$ and $k = 11$ can be explained simply as fluctuations in the environment during the runtime of the algorithm. Therefore it is hard to make a definite decision about this relationship. Again we chose $k=5$.

As observed in Figure 4 we continue further with measuring the execution times – on the *y-axis,* for different portions of the dataset. The number of input samples is placed on the *x-axis.*

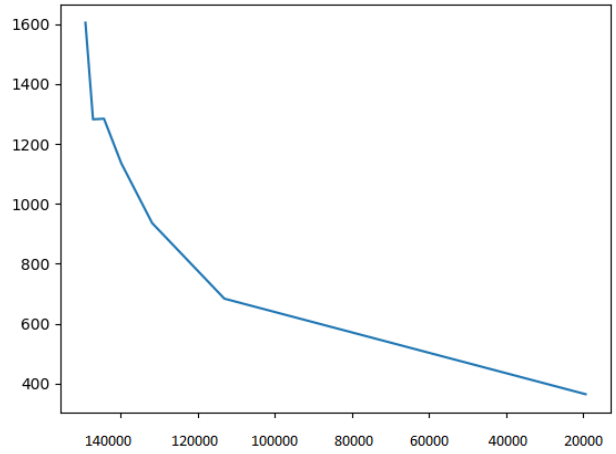

Figure 4. Evaluation time for different number of input samples

This graph confirms what we observed previously on Figure 3, but does not correspond of our expectation for the complexity of the *k-means* as discussed in Section IV. We will discuss this briefly below, but before that we need to observe the final clusters – Figure 5:

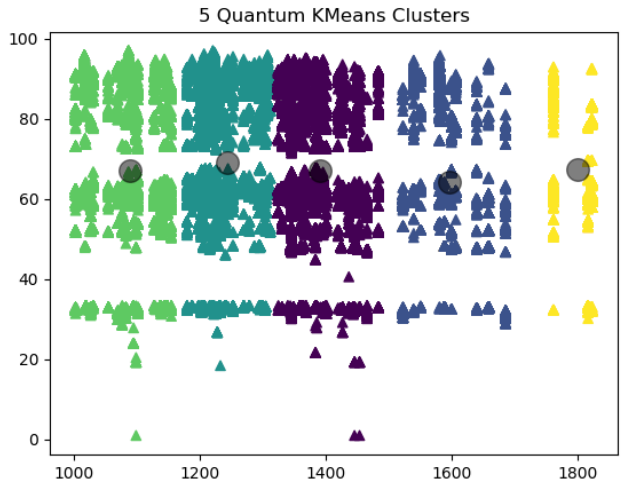

Figure 5. Final Clusters gotten by Quantum KMeans

As it can be observed the three clusters of low, middle and high randomization have been clearly separated from the most plausible real usernames to the less probable ones that have to be verified with the highest priority. The cluster centroids are marked with the bigger grey circles and we can observe that their trend is to remain in the middle of the *y-axis*.

The quantum approach provides a computational complexity of $O(INlogND)$, where I is the number of executions of the step. The *Silhouette* routine has at least $O(IN^2)$ complexity in the classical world, but for the QRAM execution we expect $O(N)$ as already described in [9]. The overall evaluation complexity would be $O(I(N + NlogND)) = O(INlogND)$. From the rules of asymptotic analysis we are removing the summation with N, but for the purposes of the paper we keep the multiplication by I. This explains the graph on Figure 4 and shows that the overall complexity is less than it would be if we used the classical approach.

VI. Conclusion

In the present paper a novel technique to reveal users' account names' validity is suggested. It uses the usernames generated by pseudo-random chaotic generator which are fed to the quantum machine learning algorithm for clustering of the names in different groups with respective probability about their truthiness. To confirm the efficiency of the proposed new technique the Silhouette method is used and the evaluation times are measured.

The chaotic pseudo-random number generator has shown very good efficiency, the binarization has been avoided thanks to the random rules attribution for each real value (the first significant digit). Moreover, the created database with fake usernames allowed to train successfully the quantum learning machine.

The results from the quantum machine learning algorithm suggested above show that this technique provides faster classification for revealing the real and the fake accounts than the other existing clustering methods. Some of more important areas for using this type of security systems are also discussed in [12, 13]. The novelty of the present technique is that it uses *quantum k-means* and *quantum Silhouette*, which are faster than the classical approaches and the results from clustering are still reliable. The quantum approach for the Silhouette method from [9] is applied. In summary, we can say that this technique can be used in important systems from the real world such as: electronic voting of government, parliament, finance institutions; e-banking; insurance; public health services; communications; continuous and discrete manufacturing etc.

Acknowledgment

The authors of the paper are thankful to the project RILA 38652 UG for supporting this research project.

And also they express their special acknowledgements to Alexandre Boissel (from High School Externat Chavagnes, Nantes, France) and to Desislav Andreev (one of the authors of this paper) for creating the program codes. First one wrote Python code to generate the fake usernames, proposed as a free software. The second one applied machine learning approach (as a theory and as a program code on Python (QuTiP) and Quantum++ (C++)) for setting the usernames in different clusters and then evaluating their quality for multiple input entries.

References

[1] ConsentCheq, "GDPR Consent Management for Small Business", US Headquarters, York, USA, European Union Office, Amsterdam, Netherland
https://www.consentcheq.com/index.php/consentiq/
(Access date: 6 February 2019)

[2] IDS Clobal Limited, DHOUND, IDS, ITBand, "Dhound Intrusion detection – security monitoring system"

https://dhound.io/

(Access date: 6 February 2019)

[3] Deutsche Gesellschaft fur Datenschutz (DGD), "DATENSCHUTZHANDBUCH DATA PROTECTION MANUAL" , Compliance with the General Data Protection Regulation (GDPR), German Association for Data Protection

https://dg-datenschutz.de/wp-content/uploads/2018/02/

DGD_Datenschutzdokumentation_Broschuere.pdf

(Access date: 6 February 2019)

[4] El Assad,H. Noura, I. Taralova, Design and Analyses of Efficient chaotic Generators for Crypto-Systems", Lecture Notes in IAENG Transactions on Electrical and Electronics Engineering, vol. I, book chapter. Publisher: IEEE Computer Society. ISBN: 978-0-7695-3555-5, 2009.

[5] R. Lozi, I. Taralova, "From chaos to randomness via geometric undersampling", European Series in Applied and Industrial Mathematics ESAIM: Proceedings and surveys, Vol. 46, 2014, pp. 177-195.

[6] O. Garasym, I. Taralova, R. Lozi, "New nonlinear CPRNG based on tent and logistic maps", Complex Systems and Networks - Dynamics, Controls and Applications, 2015, Springer pp. 131-161.

[7] O. Garasym, I. Taralova, R. Lozi, "Robust PRNG based on homogeneously distributed chaotic dynamics", Journal of Physics: Conference Series 692, 2016, 012001

http://iopscience.iop.org/1742-6596/692/1

(Access date: 6 February 2019)

[8] R. Shaw, KDnuggets, "Quantum Machine Learning: An Overview"

https://www.kdnuggets.com/2018/01/quantum-machine-learning-overview.html

(Access date: 6 February 2019)

[9] Desislav A. Andreev, "Analysis of machine learning methods and algorithms in a quantum entanglement-based environment", Proceedings of 47th sprint conference of UMB, Borovets, April 2018, pp. 129-137

[10] K. Pavan, Allam Appa Rao, A.V. Dattatreya Rao, "An Automatic Clustering Technique for Optimal Clusters", Department of Computer Applications, Rayapati Venkata Ranga Rao and Jagarlamudi Chadramouli College of Engineering, Guntur, India, Jawaharlal Nehru Technological University, Kakinada, India, Department of Statistics, Acharya Nagarjuna University, Guntur, India, September 2011.

[11] P. J. Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65, 1987, doi:10.1016/0377-0427(87)90125-7

https://en.wikipedia.org/wiki/Silhouette_(clustering)

(Access date: 6 February 2019)

[12] Andreev, D., S. Petrakieva, S., I. Taralova, "Reveal false names of accounts as a result of hackers' attacks", Proceedings of 12th International Conference for Internet Technology and Secured Transactions (ICITST – 2017), Workshop 1 Title: Chaos-based Data Protection; Data Security and Hiding in Multimedia Communications (CDP-DSHMC 2017), ISBN: 978-1-908320-79-7, 11 – 14 December 2017, University of Cambridge, UK, Published by Infonomics Society, pp. 39 – 42.

[13] Andreev, D., S. Petrakieva, I. Taralova, "A novel approach for protection of accounts' names against hackers combining cluster analysis and chaotic theory", Journal of Internet Technology and Secured Transaction' 2018, vol 7, issue 2, June 2018, pp. 579-587.