Privacy of Clients' Locations in Big Data and Cloud Computing

Imad Ali Hassoon
Faculty of Automatic and Computers
University "Politehnica" of Bucharest
Bucharest, Romania
emadali1980@yahoo.com

Nicolae Tapus
Faculty of Automatic and Computers
University "Politehnica" of Bucharest
Bucharest, Romania
ntapus@cs.pub.ro

Anwar Chitheer Jasim
Faculty of Automatic and Computers
University "Politehnica" of Bucharest
Bucharest, Romania
anwaralbazooni@gmail.com

Abstract— **Amid the very hot issues, nowadays, the one related to Locations' privacy (GPS related) finds itself in top position. When it comes to talking about clients' locations in cloud or big data, the probable risk to privacy of clients' location is one of the major challenges should to be faced. In the recent years, a lot of developers and researchers have been paying attention to improve methods to provide privacy for data of clients' locations which it always processed by third-party. Big data could be like a puzzle for many researchers if they didn't understand it in the correct-side. Big data has to be understood as the process of gathering as much data as can be permitted in order to collect knowledge out of them (ideally in ground-breaking ways). so, this concept gives us attention that is privacy of clients' locations in cloud or big data could be under risks if it is used to collect knowledge or sell it to third-party.**

In our research, we try to show how we have implemented our algorithm (Diff-Anonym) in real data set (available at http://openaddresses.io) as to offer privacy for the clients' Locations in Big Data and cloud computing, as well as to improve our previous work which was simulation in normal data that appeared little differences in the results.

Big Data; cloud; Locations' privacy; K-anonymity; differential Privacy.

I. INTRODUCTION

Big Data Location (BDL) is one of the important subjects that can be usefully and widely subjected to analysis and utilized nowadays in the computer science field. Big data location includes many options that contain in themselves the necessary resources to observe general information regarding human life and to analyze community activity.

BDL depends on geographical conditions to analyze and observe the movements of people and their activities [1].

BDL can be seen as a combination of huge human social information and geographical data that includes the identification of individuals' locations and specific times, which in its turn could, by analysis, generate new data.

Locations' privacy is a top priority when it comes to the current emergency issues that society faces. Every day the people, consciously or not, lose more and more when it comes to the privacy of their location and movements. Many organizations focus on using the locations to track their clients and provide them with information on various products.

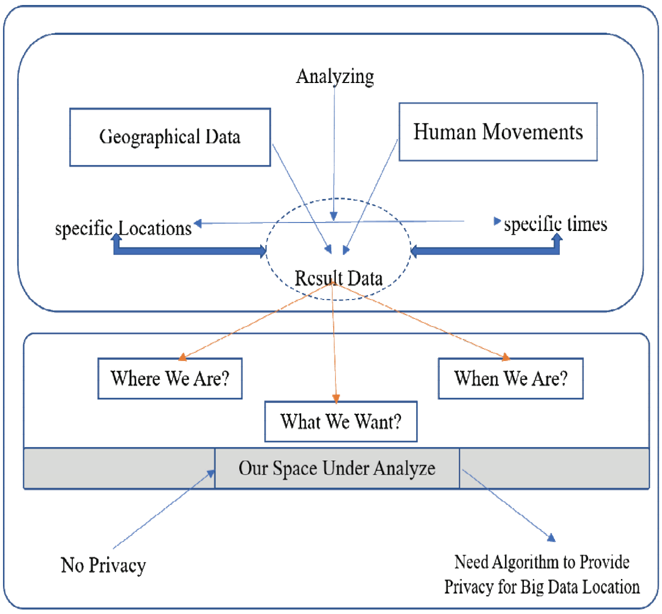


Fig 1. Illustrate Life-cycle of BDL

In fig 1 we illustrate the Life-cycle of BDL that includes the combination of human movements with geographical data that depends on (specific location – specific times), which with result data can answer the three questions: "Where are we?", "When are we?" and "What do we want?". The answers of these three questions represent an attack on our privacy from organizations that put us under analyses in order to track us or sell us goods and services.

Clients' location in big data or cloud represents geo-data (longitude and latitude) which is received from or is sent to clients about locations or nearby information. This information is collected in some systems or cloud platforms every time and after some period, when it is in huge amounts, it is considered big data. With the current trade of data between companies and all the analyses, the privacy of the clients could be under risk and subject to attacks. In a simplified way we will define some of the topics, as an introduction to our work:

Clients' location in Big Data

In recent years, Big data from different sides is used to analyze and find information about clients, such as location services, e-commerce, online gaming, advertisement services, etc. From all these IT services mentioned, we want to focus on the clients' location, in relation to location services, in order to discover the risks of analysis of Clients' location in Big Data. The important element is the privacy of the clients [1][2][3].
Big data has put too much pressure on traditional databases and structured stores of information. In reality, related to Big data, we can highlight such characteristics as velocity, volume and the variety of data, these ways having recently developed new structures to collect, analyze and generate new knowledge [4].

1. Clients' location in Cloud Computing

Cloud computing proposes solutions to manage the available resources on a pay-per-use method, the data management taking place in secure environments. Thus, the cloud computing provides services, storage, application platforms and optimized frameworks for clients. The providers of cloud computing focus on offering a flexible service, cost-effective IT infrastructure and secure environments for companies and organizations [5], [16]. Also, they manage the data of clients and analyze it as to understand what the clients want, where they stay, work, their movements, etc. The locations' services provides data about the clients' location, and the providers of cloud computing either sell or analyze this information, which can represent an attack on the privacy of the clients.

2. Privacy of Clients' location

The privacy of clients' location represents geo-data (longitude and latitude) of clients that can return information about their daily tracks, where they stay or go, the risks appearing when the clients don't allow this information to be used publicly. The variety of privacy models and which of them offers a guarantee to maintain big data privacy is an issue that requires much research and study in order to determine which is the best and the most appropriate one to be applied in the future. In this paper, we reviewed models of privacy and focused on two models to be implemented in the system for privacy management in big data.

A. K-anonymity [22].

The aim of k-anonymity is to cover data sets by restricting the intruders and not allowing disclosure of private data. Its purpose is to block any cases of individuals' identity disclosure. The objective of k-anonymity is to create a collection of quasi-identifiers in the anonymized data to indicate at least k-tuples, which are purported as equality tuples.
K-anonymity was proposed in 2002 by Sweeney [6], and was further developed in 2008 by Lodha and Thomas [7].
Definition 1: Let RT (A1, . . ., An) be a table and QIRT be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in RT [QIRT] appears with at least k occurrences in RT [QIRT] (Sweeney, 2002) [6], [8], [9].
When applying k-anonymity if k=3 or k=N, then at least three rows or N rows should have equivalence class. After achieving that, generalization is applied to the operations: first is on the dataset and after the k-anonymization technique is applied [9].

B. Differential-privacy [22].

Differential-privacy is one of the privacy models used for anonymization, which provides privacy safeguards more than other models, such as k-anonymity, T-Closeness or L-diversity [10]. Differential-privacy implies publishing the results of a query with some modifications added to the results of the query. In this case, the attacker cannot guess the results of the query because it contains a modification that has 100% guarantee of putting off the intruder [15] [16].
Nevertheless, Differential-privacy has several drawbacks. The first major impediment is that differential privacy fails to give assurances with dataset linkage and attribute in data. Usually, this model is preferred in cases where the result of congruence queries is small and with low sensitivity. This makes differential-privacy the best in restricted classes of queries. This model was initially proposed by Cynthia Dwork in 2008 [11], [12]. Finally, differential privacy as one of the most important models to provide privacy, aims to split data to small parts while adding noise to the queries to guarantee that it will not affect the analysis of the researcher nor questions the privacy of the individual [14]. Over the previous years, many ways to add noise on data to protect individual privacy have been proposedin numerous research on differential privacy [13].

II. RELATED WORKS

In 2016, TAKAHIRO HARA and all [17], came up with a model that proposed an application to which the clients send the information regarding their location, after which the application will take the respective information and convert it to anonymous information, changing the clients' location with different geographical information. Their method reduces the user's traceability by cross-users and dummies. They compared their method with VULR method and their result was 20 times better than VULR method and 5 times better than PAD method; the only exception was the Random Movement method, which was better than their methods.
Yu Wang (2016) proposed four heuristics (Algorithms 2, 3,4, 5) [18]. These algorithms are used to generate cloaking areas that further on are used to define requirements of users'

privacy. In the research Yu Wang preforms extensive simulations in 2 types of environment: the first was in real-life datasets and the second in a synthetic set. The result included several interesting observations that have been reported.
Beresford and Stajano [19] in 2003 proposed in their research a framework used for changing a user's identity through pseudonyms. In order to measure the location privacy, the research was based on two concepts: anonymity sets and entropy. Also in 2003, Gruteser and Grunwald [20], proposed a method that used the concept of k-anonymous, which was based on the fact that the user's location is reported, therefore, by applying this method, the application will guess at least k - 1 other users that are at the same time in the same location [18]. In 2016 Jingjing Wang and all [21], proposed to combine two methods: generalized k-anonymity and LPPS which uses the CRT that is designed. In their research, in order to implement their work, they depend on a trusting third-party or some other party if trust does not exist. Even they achieved a good result, but still have in their work the disadvantage of existing k-anonymity, because the k-anonymity schemes can work with exposed regions which happen during the interaction of certain nods with neighbor nodes.

III. IMPLEMENTATION OF THE PROPOSAL

In our previous work we implemented assimilation test to validate the algorithm (Diff-Anonym) [22]. The proposed work was design process of combination of the two models of privacy and implementation of the algorithm Diff-Anonym to provide privacy in normal data.

Algorithm.1 Diff-Anonym
Input: Data set from any size of data to include privacy.
Output: Data set with privacy models (k-anonymous - differential).
Step 1: Upload data into framework.
Step 2: Select fields of attributes to arrange in new temporary tables.
Step 3: Detect quasi identifier in temporary tables.
Step 4: Split tables into mini tables.
Step 5: Apply k-anonymity to mini temporary tables.
Step 6: Detect equal attributes in results.
Step 7: Spread the results of apply k-anonymity.
Step 8: Add noise to the data which already have equal attributes in results.
Step 9: Re-combine the results in big data set.

Table No 1,2 in below illustrate part of the original data set before and after of implement Diff-Anonym algorithm in previous work. We achieved our goal of reducing the possibilities of guess privacy in the case of attack on data. The advantages of previous work are increased guarantee by differential privacy and limited identity disclosure of individuals by K-anonymity method.

Table No.1 The original dataset

Sq_no	*ID_no*	*Name_ Player*	*Age_player*	*Gender_ player*	*Position_ Player*	*Status_ Player*
1	121010	Hanna	22	F	Physician assistant	Influenza
2	121011	Sami	31	M	Striker	Ready
3	121012	Lyla	23	F	Physician assistant	Malaria

Table No.2 Results of Diff-Anonym algorithm application

Sq_no	*Age_player*	*Gender_player*	*Position_Player*	*Status_Player*
1	20-25	F	Professional	Influenza
2	30+	M	Player	ready
3	>20	F	P####	Malaria

Recently we tried to examine the algorithm above with real data set to provide privacy for the clients' Locations in Big Data and cloud computing. Also, we have measured the system's time for responding to the clients' requests and what the best size is for dividing the data for interaction on the clients' side.

IV. RESULT AND DISCUSSING

The work on providing privacy for big data, the case study, includes additional information that reflected from an original dataset, which contains the locations of addresses in Bucharest (Area: 228 km² - 124802 count of recorded locations).
The dataset in case study is available at (http://openaddresses.io).

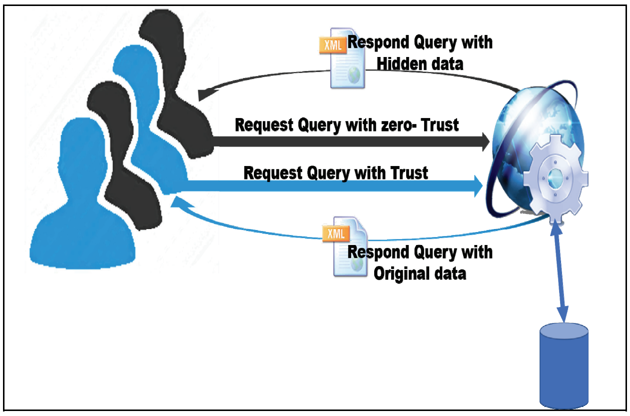


Fig 3. XML file Result with mirror reflecting Data

To keep original data in safe, we work in Diff-Anonym Algorithm to create mirror-data that reflects the respective data with privacy. We use web services to read/write data from original location and return result in Xml-file to third party location.

In figure 3,4 we have the data in XML file of respond query of clients' location. We assume in this paper to return multi results include the original fields and the fields with result of implement Diff-Anonym algorithm to compare the results and prove that proposed could work with normal data and with big data also.


```
▼<NewDataSet xmlns="">
  ▼<Table diffgr:id="Table1" msdata:rowOrder="0">
     <Long2>26.0*******</Long2>
     <Lat2>44.4*******</Lat2>
     <STREET2>Str. $$$$$$$$$</STREET2>
     <seq>26.0</seq>
     <nr>2</nr>
     <seq_b>44.4</seq_b>
     <nr_b>2</nr_b>
     <seq_c>Str.</seq_c>
     <nr_c>2</nr_c>
     <Clmov_Long>26.08553820</Clmov_Long>
     <Clmov_Lat>44.44183990</Clmov_Lat>
     <NUMBER>28</NUMBER>
     <STREET>Str. Popa Tatu</STREET>
     <DISTRICT>Sector 1</DISTRICT>
     <POSTCODE>0</POSTCODE>
     <HASH>e92a73ab11df9072</HASH>
     <Clmov_ID>2000</Clmov_ID>
   </Table>
  ▼<Table diffgr:id="Table2" msdata:rowOrder="1">
     <Long2>26.0*******</Long2>
     <Lat2>44.4*******</Lat2>
     <STREET2>Str. $$$$$$$$$$$$$</STREET2>
     <seq>26.0</seq>
     <nr>2</nr>
     <seq_b>44.4</seq_b>
     <nr_b>2</nr_b>
     <seq_c>Str.</seq_c>
     <nr_c>2</nr_c>
     <Clmov_Long>26.08114590</Clmov_Long>
     <Clmov_Lat>44.44191280</Clmov_Lat>
     <NUMBER>65</NUMBER>
     <STREET>Str. Transilvaniei</STREET>
     <DISTRICT>Sector 1</DISTRICT>
     <POSTCODE>0</POSTCODE>
     <HASH>1b2062712d154b64</HASH>
     <Clmov_ID>2001</Clmov_ID>
   </Table>
</NewDataSet>
```


Fig 4. XML file Result

Diff-Anonym Algorithm is a combination of k-anonymity and differential privacy method, it works on three levels, the first level being the one that reads the data and separates it in multi groups. At this level, we implement the algorithm in the dataset of (124802 records of locations in Bucharest) and by using multi tests we divide records into small groups (total records, half records, 10000, 5000, 1000). In the last test, we divide the entire bulk of records by 1000 records for each group of data, and discovered that this was the best result with regards to execution time and display time.

In figure 5,6 we present the results of implementing the first level of Diff-Anonym Algorithm which includes the upload of data and data division into different temporary groups. We assume the following: K = 3, the number of records in figure 5 was R = 1,000 and the time of responding on the client's side was T = 0.1083333ms. But in figure 6 we have R = 10,000 and the responding time on the clients' side was T = 0.1263333ms.

In second level of (Diff-Anonym algorithm) we implement k-anonymity in groups of data return from first level which include fields of Clients movements (longitude and latitude) with range of k between (1-9). The implemented return result test with various context and the result in range <5 was more benefits to support privacy of records.

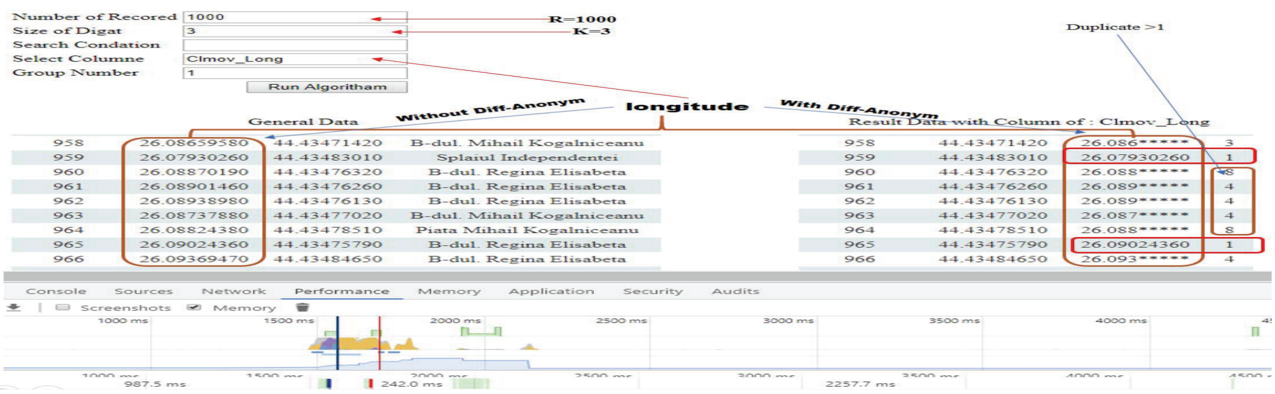


Fig 5. Illustrate implement R=1000 K=3 T= 0.1083333m

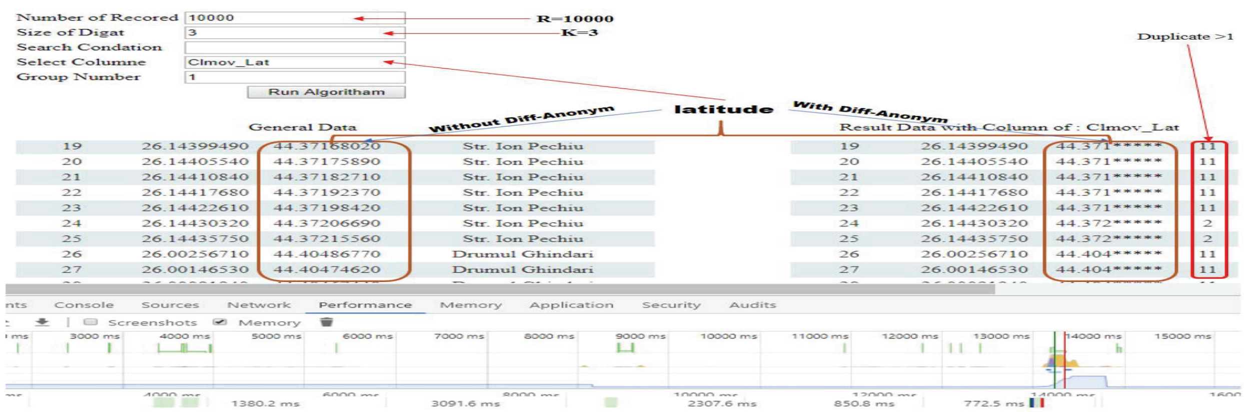
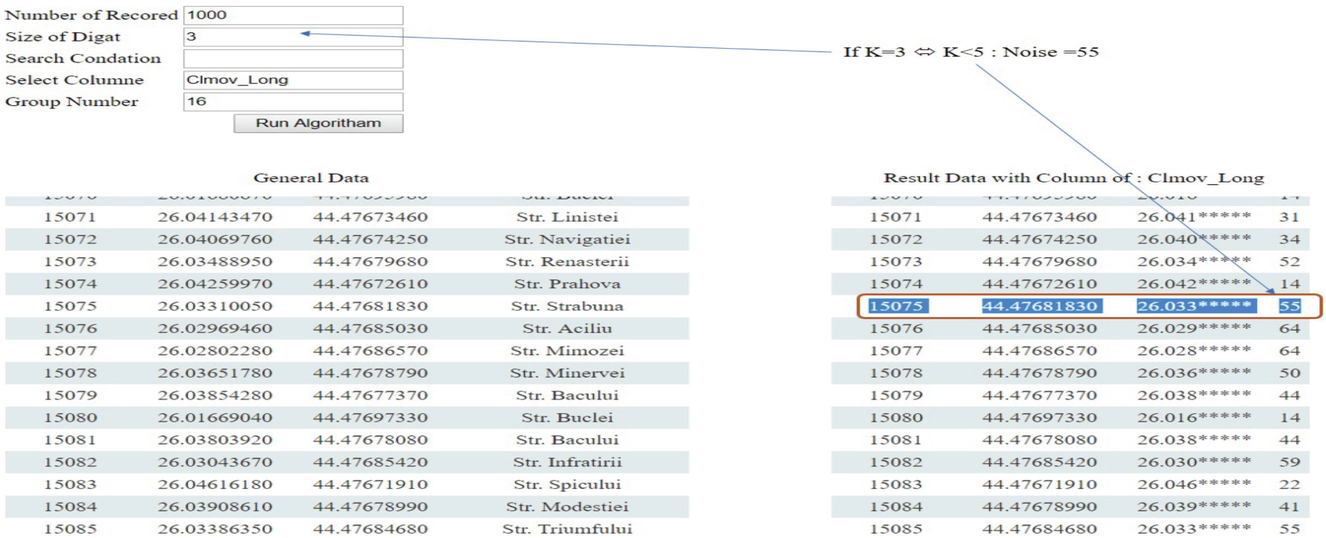


Fig 6. Illustrate implement R=10000 K=3 T= 0.1263333m

In figure 7,8 we present the results of implementing the second level of Diff-Anonym Algorithm which includes the test of the result when K grater or smaller than 5. In figure 7 we implemented K = 3, which means it is smaller than 5, and we get the result with 55 records for ID = 15,076 but when we assume K = 6, so greater than 5, the result was 2 records for the same ID = 15,076.

Third level in (Diff-Anonym algorithm) we re-read the result and discover the records that still have similarity with other records and implement differential methods to cover the records that have similar tuples in the results.

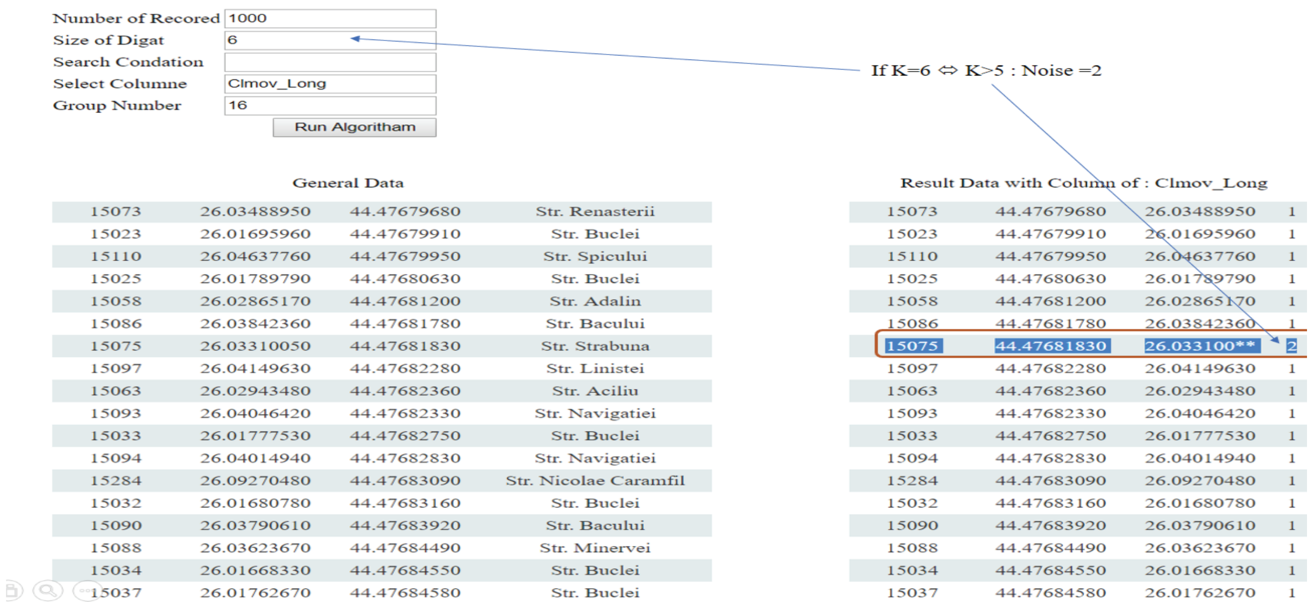
Differential-privacy implies publishing the results of a query with some modification added to the results of the query. In this case, the attacker cannot guess the results of the query because it contains a modification that has 100% guarantee of putting off the intruder.



15071	26.04143470	44.47673460	Str. Linistei
15072	26.04069760	44.47674250	Str. Navigatiei
15073	26.03488950	44.47679680	Str. Renasterii
15074	26.04259970	44.47672610	Str. Prahova
15075	26.03310050	44.47681830	Str. Strabuna
15076	26.02969460	44.47685030	Str. Aciliu
15077	26.02802280	44.47686570	Str. Mimozei
15078	26.03651780	44.47678790	Str. Minervei
15079	26.03854280	44.47677370	Str. Bacului
15080	26.01669040	44.47697330	Str. Buclei
15081	26.03803920	44.47678080	Str. Bacului
15082	26.03043670	44.47685420	Str. Infratirii
15083	26.04616180	44.47671910	Str. Spicului
15084	26.03908610	44.47678990	Str. Modestiei
15085	26.03386350	44.47684680	Str. Triumfului

15071	44.47673460	26.041*****	31
15072	44.47674250	26.040*****	34
15073	44.47679680	26.034*****	52
15074	44.47672610	26.042*****	14
15075	44.47681830	26.033*****	55
15076	44.47685030	26.029*****	64
15077	44.47686570	26.028*****	64
15078	44.47678790	26.036*****	50
15079	44.47677370	26.038*****	44
15080	44.47697330	26.016*****	14
15081	44.47678080	26.038*****	44
15082	44.47685420	26.030*****	59
15083	44.47671910	26.046*****	22
15084	44.47678990	26.039*****	41
15085	44.47684680	26.033*****	55

Fig 7. Illustrate implement R=1000 K=3=> K<5 N=55



15073	26.03488950	44.47679680	Str. Renasterii
15023	26.01695960	44.47679910	Str. Buclei
15110	26.04637760	44.47679950	Str. Spicului
15025	26.01789790	44.47680630	Str. Buclei
15058	26.02865170	44.47681200	Str. Adalin
15086	26.03842360	44.47681780	Str. Bacului
15075	26.03310050	44.47681830	Str. Strabuna
15097	26.04149630	44.47682280	Str. Linistei
15063	26.02943480	44.47682360	Str. Aciliu
15093	26.04046420	44.47682330	Str. Navigatiei
15033	26.01777530	44.47682750	Str. Buclei
15094	26.04014940	44.47682830	Str. Navigatiei
15284	26.09270480	44.47683090	Str. Nicolae Caramfil
15032	26.01680780	44.47683160	Str. Buclei
15090	26.03790610	44.47683920	Str. Bacului
15088	26.03623670	44.47684490	Str. Minervei
15034	26.01668330	44.47684550	Str. Buclei
15037	26.01762670	44.47684580	Str. Buclei

15073	44.47679680	26.03488950	1
15023	44.47679910	26.01695960	1
15110	44.47679950	26.04637760	1
15025	44.47680630	26.01789790	1
15058	44.47681200	26.02865170	1
15086	44.47681780	26.03842360	1
15075	44.47681830	26.033100**	2
15097	44.47682280	26.04149630	1
15063	44.47682360	26.02943480	1
15093	44.47682330	26.04046420	1
15033	44.47682750	26.01777530	1
15094	44.47682830	26.04014940	1
15284	44.47683090	26.09270480	1
15032	44.47683160	26.01680780	1
15090	44.47683920	26.03790610	1
15088	44.47684490	26.03623670	1
15034	44.47684550	26.01668330	1
15037	44.47684580	26.01762670	1

Fig 8 Illustrate implement R=1000 K=6=>K>5 N=2

V. Conclusion and Future Work

Many organizations focus on using the locations of people in order to track the movements of their clients and thus offer them various products. In the recent years, a lot of researchers and developers have been paying attention to these issues, especially in big data and cloud computing, because these two types include huge data of the locations of individuals, which will further be used by the third party. In our research, we presented the results of implementing the Diff-Anonym Algorithm to provide privacy for clients' locations. For this purpose, we have tested our algorithm on data set that contains addresses locations in Bucharest (Area: 228 km² - 124802 count of locations). The dataset of our case study is available at (http://openaddresses.io).

We have built our structure based on a non-trusted third party. This structure gives us the flexibly to send the data with a certain scheme depending on who sent the request to read our data. Also, with privacy we keep our original data in safety and manage the allow/deny function with Admin Side in order to hide or display data cautiously and therefore not lose owner data or the privacy of clients' locations. In the future, we will focus to improve our algorithm and combine all models in a framework that will be tested with real data and executed in real time. In addition, we will compare our result with other frameworks in big data or cloud computing.

I. References

[1] Dong, SH., Zhang, HW., Zhang, LB. et al. Pet. Sci. (2017) Use of community mobile phone big location data to recognize unusual patterns close to a pipeline which may indicate unauthorized activities and possible risk of damage . https://doi.org/10.1007/s12182-017-0160-7

[2] Daggitt ML, Noulas A, Shaw B, et al. (2016) Tracking urban activity growth globally with big location data. R Soc Open Sci. 2016;3(4):150688. doi:10.1098/rsos.150688.

[3] Liu JN. (2012) The recent progress on high precision applications of Beidou navigation satellite system. Report of the stanford's 2012 PNT challenges and opportunities symp. (SCPNT 2012), 2012. (in Chinese).

[4] Soria-Comas, J. & Domingo-Ferrer, J. Data Sci. Eng. (2016) Big Data Privacy: Challenges to Privacy Principles and Models. https://doi.org/10.1007/s41019-015-0001-x

[5] Jakimoski, Kire. (2016). Security Techniques for Data Protection in Cloud Computing. International Journal of Grid and Distributed Computing. 9. 49-56. 10.14257/ijgdc.2016.9.1.05.

[6] Sweeney L .(2002) "k-anonymity: a model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, No.5 , pp.557–570.

[7] Lodha, S. and Thomas, D. (2008) "Probabilistic anonymity", in Privacy, Security, and Trust in KDD, pp.56–79, Springer, Berlin, Heidelberg.

[8] Jordi Soria-Comas, Josep Domingo-Ferrer, (2016),"Big data privacy: challenges to privacy principles and models", Data Sci. Eng. (2016) 1(1):21–28, DOI 10.1007/s41019-015-0001-x

[9] Nancy Victor* and Daphne Lopez, (2016), "Privacy models for big data: a survey", January 2016, DOI: 10.1504/IJBDI.2016.073904.

[10] Sánchez D., Domingo-Ferrer J., Martínez S. (2014), "Improving the Utility of Differential Privacy via Univariate Microaggregation", Domingo-Ferrer J. (eds) Privacy in Statistical Databases. PSD 2014. Lecture Notes in Computer Science, vol 8744. Springer, Cham.

[11] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum and S. Vadhan (2009). "On the complexity of differentially private data release: efficient algorithms and hardness results". In: Proc. of the 41st Annual Symposium on the Theory of Computing-STOC 2009, pp. 381-390, 2009.

[12] C. Dwork, (2008) "Differential privacy: a survey of results", in Theory and Applications of Models of Computation, pp.1–19, Springer, Berlin, Heidelberg.

[13] McSherry, F. and Talwar, K. (2007) 'Mechanism design via differential privacy', in 48th Annual IEEE Symposium on Foundations of Computer Science, 2007, FOCS'07, IEEE, pp.94–103.

[14] Li, C. et al. (2010) 'Optimizing linear counting queries under differential privacy', Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, ACM.

[15] Friedman, A. and Schuster, A. (2010) 'Data mining with differential privacy', in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp.493–502.

[16] Xiao, Y., Xiong, L., Yuan, C. (2012). "Differentially private data release through multidimensional partitioning". Proceedings of the 7th VLDB conference on Secure data management (SDM'10), pp. 150-168 (2010).

[17] T. Hara, A. Suzuki, M. Iwata, Y. Arase and X. Xie,(2016) "Dummy-Based User Location Anonymization Under Real-World Constraints," in IEEE Access, vol. 4, pp. 673-687, 2016. doi: 10.1109/ACCESS.2016.2526060.

[18] Y. Wang, D. Xu and F. Li, (2016) "Providing location-aware location privacy protection for mobile location-based services," in Tsinghua Science and Technology, vol. 21, no. 3, pp. 243-259, June 2016. doi: 10.1109/TST.2016.7488736.

[19] A. R. Beresford and F. Stajano, Location privacy in pervasive computing, IEEE Pervasive Computing, vol. 2, pp. 46–55, 2003.

[20] M. Gruteser and D. Grunwald, Anonymous usage of location-based services through spatial and temporal cloaking, in Proc. of ACM MobiSys, 2003.

[21] Wang, Jingjing & Han, Yiliang & Yang, Xiaoyuan. (2016). An Efficient Location Privacy Protection Scheme Based on the Chinese Remainder Theorem. Tsinghua Science and Technology. 21. 260-269. 10.1109/TST.2016.7488737.

[22] Imad. Ali. Hassoon, N. Tapus and Anwar. C. Jasim,(2017) "Enhance privacy in big data and cloud via diff-anonym algorithm," 2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet), Tg. Mures, Romania, 2017.