IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol-6, Special Issue-6, July 2018 E-ISSN: 2347-2693

Trajectory Anonymization Through Generalization of Significant Location Points

Rajesh N[1*], Sajimon Abraham[2], Shyni S Das[3]

[1]Research Scholar, School of Computer Sciences, M.G. University, Kottayam, India
[2]Dept. of IT, School of Management and Business Studies, M.G. University, Kottayam, India
[3]Dept. of Computer Applications, S.A.S. S.N.D.P. Yogam College, Konni, Pathanamthitta, India

[*]*Corresponding Author: rajeshshyni2000@gmail.com, Tel.: +91-9447116541*

Available online at: www.ijcseonline.org

***Abstract*—** The widespread use of Location Based Systems results in the accumulation of movement trajectory details in a massive scale. These mobility traces are very much useful for the researchers and the developers who needs to develop or invent new mobility management applications or modify the existing ones. But without proper privacy preserving mechanism for the published trajectory details may definitely raises the issue of privacy breach for the user. So before publishing the trajectory details suitable anonymization approach has to be applied. It is also found that the protection of significant points is better than the unnecessary anonymiztion whole trajectory points. This paper proposes a new model, which depicts a model that safeguards the significant points from the malevolent attacks by the help of generalization approach. With this model, the significant location points are hided in a specified size diversified area zone. The analysis shows that this approach is well ahead of the similar approaches used by the researches and provides better privacy and less information loss.

***Keywords*—**Anonymization, Trajectory Publication, Privacy Preservation

I. INTRODUCTION

With the ubiquity of location based systems(LBS) allows users to exchange their mobility traces and they can use variety of location based services like social media, GPS(Global Positioning System) based navigation systems etc. The researchers also used these trajectory details to make or modify new or existing applications and the government authorities used it for the applications like traffic management etc. But sharing trajectory details, especially location details may reveal their social customs, habits, health details, religious customs etc. This is going to be the major privacy threat to the user.

The Privacy preservation of trajectory means to safeguard the entire trajectory from re-identification so that to overcome the disclosure of significant or important location points. In order to keep away from adversaries, we have to find the stay points[6] from the mobility traces and differentiate them as significant and non-significant points and anonymize these points is well enough for the protection of entire trajectory from malevolent attacks.

In this work, we mainly focused to anonymize the trajectory before the publication of trajectory details to stakeholders. For this purpose, it is necessary to extract halting points and identify the significant and non-significant location points from the user trajectories. At last during the anonimization of these points is to be done with the help of generalized area zone, which contains a user specified number of significant and non-significant points.

The organization of the paper is organized as follows, Section I contains the introduction about the privacy preserved data publishing and the brief summary of the work done, Section II contains the related work of trajectory anonymization and the developments in the said area, Section III contains the basic terminologies used in the paper as problem definitions, Section IV contains the proposed work and which is categorized as block diagram of the work, methodologies and algorithm which we used for the proposed work, Section V explains the results and discussion through sample screens and analysis graphs and finally Section VI concludes research work with future directions.

II. RELATED WORK

For the development and modification of new or existing trajectory management application must need published trajectory details and it is too good to have this in a privacy preserved mode. In the privacy preserved data publishing scenario, a plenty of approaches and models were there. Some of them are discussed below.

The concept of *k*-anonymity was introduced in [2] to protect health details from malevolent attack so that it is indistinguishable from *k-1* other records. Few years later, the authors in [4] introduced another model called *t*-closeness,

which states that the distribution distance between two significant attributes must not be greater than a disclosure threshold *t*. The *t*-closeness model was actually the refinement of the concept called *l*-diversity [3], which defines that each equivalence class has at least *l*-well represented diverse location values. Later an approach in [1] specifies another model called k^m-anonymity model, which uses Euclidean distance during the transaction data publishing phase and it will restricts the chance of identity disclosure. This method is also not appropriate for the generalization approaches during the distance based calculations.

III. PROBLEM DEFINTIONS

Let us consider an individual's trajectory T, which comprises of different location points and is defined as $T=\{tr_{id}, (lat_1, lon_1, tm_1), (lat_2, lon_2, tm_2), \ldots\ldots\ldots\ldots, (lat_n, lon_n, tm_n)$, where tr_{id} is the id of the trajectory and location coordinates of the moving object is (lat_j, lon_j). The tm_j is the sampling time temporal factor in a spatio –temporal coordinate (lat_j, lon_j, tm_j) of the j^{th} location.

A. *Halting points*

A halting point is a location point where the user has stayed for a particular threshold time λt and which is represented as $(Hpt_{id}, Hpt_t, lat_j, lon_j, \lambda t)$, where Hpt_{id} is the identifier of the halting point, Hpt_t is the halting point duration and (lat_J, lon_j) is the j^{th} location coordinate of the halt. The halting points can be either significant or non-significant halting points.

B. *Significant location points*

Significant location points are the halting location points where the user has some important information there. This can considered by taking the halting points which are greater or equal to a time threshold Δt.

C. *Non-significant location points*

Non-significant location points are halting location points which are not significant location points obviously which is less than the time threshold Δt.

D. *Area zone*

An area zone AZ, consists of n number of significant and non-significant halting points. The AZ is represented as $(AZ_{id}, TL_c, LR_c, Sl_n, NSl_n)$, where AZ_{id} is the area zone identifier ; TL_c and LR_c represents the top left and lower right corner coordinates of the area zone; and Sl_n and NSl_n represents number of significant and non-significant location points included in the area zone.

IV. METHODOLOGY

A. *Proposed approach*

The main work of this paper is the anonymization of original trajectories from the trajectory database (SptDb) and release the anonymized trajectory database (PubDb) for the stakeholders. The trajectory anonymization consists of the processes like halting point extraction, significant point identification and finally the area zone creation.

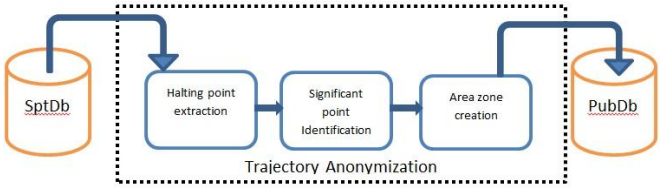


Figure 1. Block diagram of our approach

B. *Halting point extraction*

For the extraction of halting points we choose the methodology as in [7] with minor alterations. The location points where the user has stayed at least for μt is considered as halting points. i.e, for a trajectory $T = \{(Ln_1, t_1), (Ln_2, t_2), \ldots\ldots\ldots, (Ln_n, t_n)\}$, where Ln_j represents the coordinate of the location point at the sampling time t_i. But if $|ti+1 – ti| > \lambda t$, where λt is the time threshold specified by the user, then Ln_i is treated as halting point.

C. *Significant point identification*

We have identified the significant and non-significant points from the halting points and the method adopted from [1] with slight modifications. The identification of significant points includes sorting of halting points and we choose them with the strategy that by fixing a threshold time Δt. The significant points which satisfies this Δt or more and for the non-significant which are less than Δt.

D. *Area zone creation*

The area zone is a rectangular area on the trajectory, which is having user specified number of significant and non-significant halting points. When we increase the size of the area zone, the privacy increases but the information loss also increases. The Haversine distance measure is used for the distance calculation between halting points.

Haversine distance formula =

*2*6371000*asin (sqrt((sin ((lat2*x- lat1*x))/2)^2 + cos(lat2*x)*cos(lat1*x)*sin(((lon2*x- lon1 *x)/2))^2))* (1)

Where x=3.14159/180 and lat and lon are latitude and longitude respectively.

To calculate the information loss measure, we used the formula as in [5]

$$InfL = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}(1 - 1/area\_size\_of\_Zone(AZ_i,t_j))+ \sum_{x=1}^{k} L_x\right)/(n \times m) \quad (2)$$

Here *infL* represents average information loss happened at the area zone AZ_i at time ti where AZ_i stayed. L_x is the significant point that we are replacing with area zone. *n x m* denotes the total number of location points. Ultimately, the *infL* lies between 0 and 1, i.e. 0<*infL*<1

Algorithm : Generalization of Significant trajectory halting points

Input : Spatio-temporal Trajectory points from *SptDb*
Output : Anonymized trajectory points and other location points to *PubDb* for publication

1. Pre-process & Read spatio-temporal points from *SptDb* to *SpTrj*
2. Initialize *Hlt_tab1, S_tab2, NS_tab3* < -- φ
3. Set halting threshold time *th_t* < -- *μt*
4. *while (!eof(SpTrj))*
   a) *for (i=0, k=1 up to n) do*
      (i) Calculate halting point λt = | t_{i+1} - t_i |
      (ii) If (*λt* >= *μt*) then store
      Hp_k -->*Hlt_tab1, i++, k++*
5. Set significant threshold time S_t < -- Δt
6. *while (!eof(Hlt_tab1))*
   a) *for (i=0 to n) do*
      (i) if (*λt* (Hp_{ki}) > *Δt*) then
         store Hp_i -- > *S_tab2* else
         store Hp_i -- > *NS_tab3, i++*
      (ii) calculate Haversine distance *Hdt* from previous halt.
7. Create area zone AZ_i with user specified location points from *S_tab2* and *NS_tab3* using generalization technique
8. Store the TL_c and LR_c of AZ_i and publish these coordinates instead of significant points along with other location points

V. RESULTS AND DISCUSSION

We did our experiments using the real trajectory dataset available in the internet from the Microsoft's Geolife project [8], which includes the 5 year trajectories of 182 users mainly in Beijing area. The experiments were conducted on a windows 10 computer with 4GB RAM and Intel i5-3337U CPU @1.80GHz Processor. Here are some sample screens

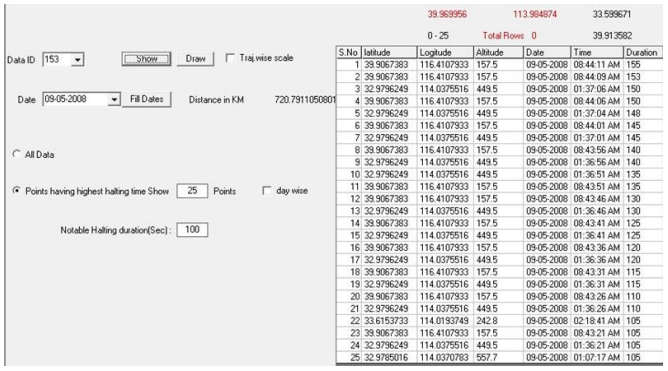
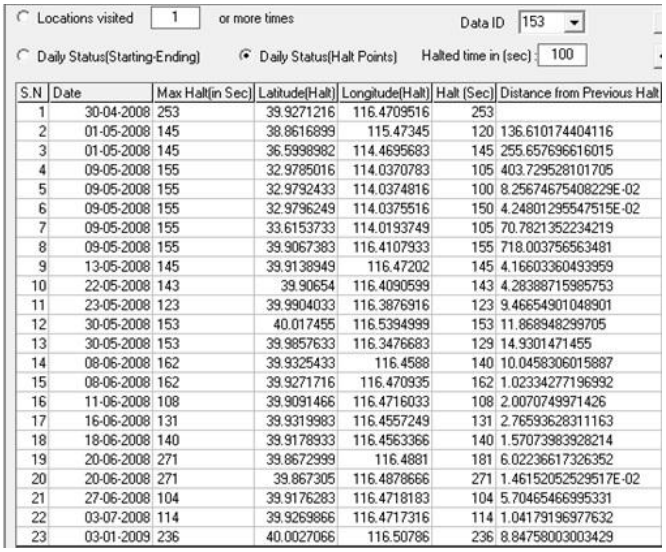


Figure 2. The first 25 halting points on a single day



S.N	Date	Max Halt(in Sec)	Latitude(Halt)	Longitude(Halt)	Halt (Sec)	Distance from Previous Halt
1	30-04-2008	253	39.9271216	116.4709516	253	
2	01-05-2008	145	38.8616899	115.47345	120	136.610174404116
3	01-05-2008	145	36.5998982	114.4695683	145	255.657696616015
4	09-05-2008	155	32.9785016	114.0370783	105	403.729528101705
5	09-05-2008	155	32.9792433	114.0374816	100	8.256746754088229E-02
6	09-05-2008	155	32.9796249	114.0375516	150	4.248012955475515E-02
7	09-05-2008	155	33.6153733	114.0193749	105	70.7821352234219
8	09-05-2008	155	39.9067383	116.4107933	155	718.003756563481
9	13-05-2008	145	39.9138949	116.47202	145	4.166033604939 59
10	22-05-2008	143	39.90654	116.4090599	143	4.28388715985753
11	23-05-2008	123	39.9904033	116.3876916	123	9.46654901048901
12	30-05-2008	153	40.017455	116.5394999	153	11.868948299705
13	30-05-2008	153	39.9857633	116.3476683	129	14.9301471455
14	08-06-2008	162	39.9325433	116.4588	140	10.0458306015887
15	08-06-2008	162	39.9271716	116.470935	162	1.02334277196992
16	11-06-2008	108	39.9091466	116.4716033	108	2.0070749971426
17	16-06-2008	131	39.9319983	116.4557249	131	2.76593628311163
18	18-06-2008	140	39.9178933	116.4563366	140	1.57073983928214
19	20-06-2008	271	39.8672999	116.4881	181	6.02236617326352
20	20-06-2008	271	39.867305	116.4878666	271	1.461520525295 17E-02
21	27-06-2008	104	39.9176283	116.4718183	104	5.70465466995331
22	03-07-2008	114	39.9269866	116.4717316	114	1.04179196977632
23	03-01-2009	236	40.0027066	116.50786	236	8.84758003003429

Figure 3. The Haversine distance from previous halt for a user


Figure 4. The connecting graph between significant points

According to our algorithm, we took trajectory traces of a single user, preprocess it and extracted the halting points. The Figure 2 shows the first 25 highest halting points on a single day for a user. The Geolife trajectory dataset was prepared with 6 digit floating point accuracy; we got 23 halting points with a halting threshold λt with 100seconds as shown in Figure 3. From this, we took some highest duration points as significant points with Δt as 140seconds. The Figure 4 shows the connecting graph between significant points. We also measures the Haversine distance from the

nearest halting points. The balance halting points were considered as non-significant points. Then we created area zone with the varying no of significant and non-significant points.

E. Evaluation

For the evaluation, we took two analyses. First one as shown in Figure 5, is for analyzing area zone creation time, and the next one Figure 6, is for evaluating the information loss measure

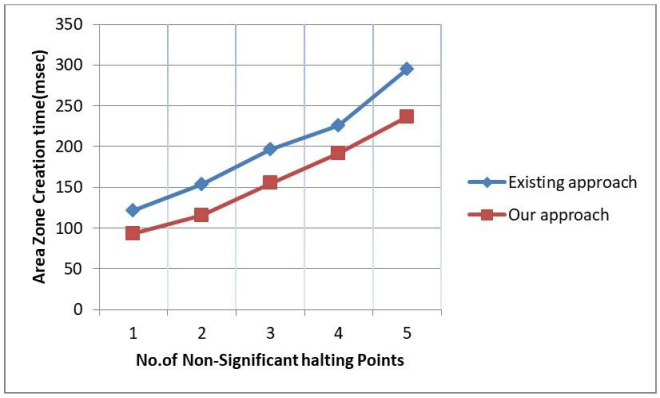


Figure 5. Analysis 1 : Area zone creation time Vs No.of non-significant points

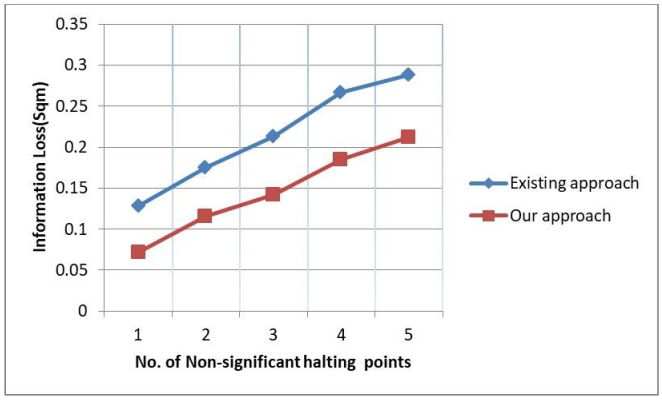


Figure 6. Analysis 2 : Information loss Vs No. of non-significant points

The analysis 1 shows that the processing time taken to create area zone with our approach is better than the existing approach in [6]. In the analysis 2 shows that information loss in our approach is less than existing approach in [1], but there is a fact that whenever we tries to increase the privacy of the user the information loss will also increases. We use Haversine distance measure due to the spherical shape of the earth and found that this one has more accuracy than the Euclidean approach.

VI. CONCLUSION AND FUTURE SCOPE

The privacy is the major concern of every individual, but at the same time avoiding LBS facilities is too hard in the digital era. So accumulation and publication of mobility traces needs to be done with most care. This paper proposes and analyses the anonymization mechanism through generalization of significant points, shows that it is relevant and suffer only a little information loss than the existing approaches. In future, we extend this approach to overcome the various malevolent linkage attacks.

REFERENCES

[1] Poulis, G., Skiadopoulos, S., Loukides, G., Gkoulalas-Divanis, A.: Apriori-based algorithms for k^m-anonymizing trajectory data. Transactions on data privacy, 7:2, pp. 165-194 (2014)

[2] Sweeney, L.: Achieving *k*-anonymity privacy protection using generalization and suppression. International Journal of uncertainty, fuzziness and knowledge-based systems, 10(5), pp. 571-588, 2002.

[3] Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: *l*-diversity: Privacy beyond *k*-anonymity. TKDD, 1(1), 2007.

[4] Li, N., Li, T., Venkatasubramanian, S.: *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity. In ICDE, pp. 106-11, 2007.

[5] Yarovoy, R., Bonchi, F., Lakshmanan, S., Wang, W.H.: Anonymizing moving objects: How to hide a MOB in a crowd? In:12th Int. Conf. on extending database technology, pp. 72-83, ACM press, New York ,2009.

[6] Huo, Z., Meng, X., Hu, H.,Huang, Y.: You can walk alone: Trajectory privacy preserving through significant stays protection. *DASFAA 2012*, Part 1, LNCS 7238, pp. 351-366, Springer-Verlag Berlin Heidelberg, 2012.

[7] Zheng, Y., Zhang, L., Xie, X., Ma, W.: Mining interesting locations and travel sequences from GPS trajectories. In: 18th International conference on World Wide Web, pp.791-800, ACM press, New York , 2009

[8] Microsoft Research Geolife, *http://research.microsoft.com/en-us/projects/geolife/*

[9] Rajesh N, Sajimon Abraham, "Privacy preserved approach for trajectory anonymization through the zone creation for halting points", International conference on Networls and Advances in Computational Technologies (NetACT17), IEEE explore, pp.229-234, 2017

Authors Profile

Mr. Rajesh N (MCA). He is a Research Scholar in the School of Computer Sciences, M. G. University, Kottayam, Kerala, India and also working as an Assistant Professor in the Department of Computer Applications, S.A.S. S. N. D. P. Yogam College, Konni, Pathanamthitta, Kerala, India. He has 18 years of undergraduate and 8 years of post-graduate teaching experience. His main area of research interests are Privacy computing and trajectory data publishing, Big data Analysis and Data Mining. He has published more than 7 articles in International and National journals and Conference proceedings.

.

Dr. Sajimon Abraham. (MCA, MSc. (Mathematics), MBA, PhD (Computer Science). He has been working as Faculty Member in Computer Applications & IT, School of Management and Business Studies, Mahatma Gandhi University, Kottayam, Kerala, India. He currently holds the additional charge of Director (Hon), University Center for International Co-operation. He was previously working as Systems Analyst in Institute of Human Resource Development, Faculty member of Computer Applications in Marian College, Kuttikkanam and Database Architect in Royal University of Bhutan under Colombo Plan on deputation through Ministry of External Affairs, Govt. of India. His research area includes Data Science, Spatio-Temporal Databases, Mobility Mining, Sentiment Analysis, Big Data Analytics and E-learning and has published 52 articles in National, International Journals and Conference Proceedings.

Shyni S. Das, M.C.A. She is currently working as an Assistant Professor in the Department of Computer Applications, S. A. S. S. N. D. P. Yogam College, Konni, Pathanamthitta, Kerala, India. She has 18 years of undergraduate and 8 years of post-graduate teaching experience. Her main area of research interests are Privacy computing, Internet of Things and Data Mining. She has published and presented more than 3 articles in International and National Journals and Conference proceedings.

