A Survey on Location Based Services in Data Mining

Ipsa Das, Md Imran Alam, Jayanti Dansana

Abstract*-- Data privacy has been the primary concern since the distributed database came into the picture. More than two parties have to compile their data for data mining process without revealing to the other parties. Continuous advancement in mobile networks and positioning technologies have created a strong challenge for location-based applications. Challenges resembling location-aware emergency response, location-based advertisement, and location-based entertainment. Privacy protection in pervasive environments has attracted great interests in recent years. Two kinds of privacy issues, location privacy and query privacy, are threatening the security of the users. The novel combined clustering algorithm for protecting location privacy and query privacy, namely ECC, is discussed. ECC applies an iterative K-means clustering method to group the user requests into clusters for providing location safety while utilizing a hierarchical clustering method for preserving the query privacy.

***Index Terms*-- Location Based Services (LBSs), K-Anonymity, Location K-Anonymity, Clustering, Clustering Cloak**

I. INTRODUCTION

Recently, Location based services (LBSs) have been originally customized for mobile users. Location based services (LBSs) are accepted to form an imperative part of the future computing surroundings that will be effortless and universally integrated into our lives. Location based services are one of the most enviable classes of services to be offered in pervasive computing environments. Service provider envisages offering many new services based on a user's location as well as augmenting many existing services with location information. Location based services are widely used in our day to day life. So user required to maintain privacy.

Privacy in pervasive computing environments includes anonymity, context, confidentiality and integrity. Location Privacy is a particular type of context privacy. It is defined as the ability to prevent other unauthorized parties from learning one's current or past location. In LBSs, there are two types of location privacy [5]: Personal Subscriber Level Privacy and Corporate Enterprise Level Privacy. Personal subscriber-level privacy must supply rights and options to individuals to control when, why, and how their location is used by an application, and to prevent other parties from learning one's past or current location. Location privacy threats refer to the risks that an adversary can obtain the mobile user location data.

Furthermore, if the LBS provider is unreliable, the location information may be abused and the users may face undesired advertisements, e-coupons, etc. Motivated by this fact, a new method of protecting location privacy based on clustering is developed.

Manuscript Received on May, 2014.

Ipsa Das, M.Tech research scholar, School of Computer Engineering, KIIT University, Bhubaneswar, India.

Md. Imran Alam, M.Tech research scholar, School of Computer Engineering, KIIT University, Bhubaneswar, India.

Jayanti Dansana, Asst. Prof., School of Computer Engineering, KIIT University, Bhubaneswar, India.

Specially, we thwart an attacker from inferring the real location information of the mobile user by adapting the K-anonymity technique to the spatial domain.

The concept of K-anonymity was introduced as characterizing the degree of data protection with respect to inference by linking and K-anonymity can be ensured in information release by generalizing and/or suppressing part of the data to be revealed. A data release is said to meet K-anonymity if every topple released cannot be related to less than K respondents, where K is a positive integer set by the data holder. In order to defend the location information of mobile users in the context of LBSs, Gruteser and Grunwald [1] firstly employed K-anonymity [1]. A subject is deemed as K-anonymity with respect to location information, if and only if the location information sent from one mobile user is indistinguishable from the location information of at least K-1 other mobile users. The spatio-temporal cloaking assumes that all users have the same K-anonymity requirements and K cannot vary with the different privacy requirements of different users. In order to increase the scalability, a customizable K-anonymity model instead of a uniform K was proposed in. Every user can specify a different K-anonymity value based on his minimum anonymity level and his favoured spatial and temporal tolerance level in order to maintain the personalized variable privacy requirements.

Location based services are widely used in our day to day life. So user required to maintain privacy. Two kinds of privacy concerns should be paid close attention to: 1) Location Privacy and 2) Query Privacy. Location privacy emphasizes on preventing the disclosure of location information collected by the adversary from being used to identify a service user. Query privacy is an additional protection for the location privacy preservation. It focuses on preventing the attacker from figuring out the identity of the user through analyzing the contents of the query. In that case, although the location privacy is well-protected, the attacker can still successfully link the identity of the users to the queries by referring to the background knowledge of the users he owns.

II. ARCHITECTURE OVERVIEW OF LBS

The communication is done by a mobile client through the anonymity server with the third party LBS provider. It is a secure gateway to use the anonymity server to the semi honest LBS provider for the mobile client. It runs the message perturbation engine, then location perturbation is performed on the message received from the mobile clients it is forwarded to the LBS provider. The message contains the location information of the mobile client and a time stamp in addition to service-specific information, which is provided to the LBS provider.

After the message has been received from a mobile client, then any identities such as Internet Protocol (IP) addresses, perturbs the location information through spatio-temporal cloaking are removed by the anonymity server and then forwards the anonymized message to the LBS provider. It is referred by the spatial range for replacement of a 2D point location lies anywhere within the range. Temporal cloaking is the replacement of a time point associated with the location point with the time interval which includes the original time point.

System architecture consists of trusted third parity which is acting as a middle layer between mobile user and LBS provider. First the exact location information is received by TTP from a mobile user then exact information will blured by TTP in to the cloaked spatial area using a clustering algorithm. A list of results will be sent to TTP because the LBS provider cannot receive the exact location but the cloaked area. Finally, most optimal results are selected by TTP to the mobile user from the list therefore LBS are enjoyed by the mobile user and get more privacy without revealing his private location information. The following six processes describe the whole process and illustrated in the fig 1.

- Every mobile user sends a message consisting of his exact location, K, and an LBS request.
- All users are clustered as soon as TTP receives the message.
- The exact location information is replaced by MBR of the cluster where the mobile user locates.
- LBS return a list of results to TTP in light of the received MBR from step 3.
- TTP sends the optimal result to the mobile user based on the exact location information in step 1.
- The mobile user receives the result from step 5. Since a mobile device possesses limited memory and limited computing capabilities, it cannot act as an anonymity server instead of TTP.

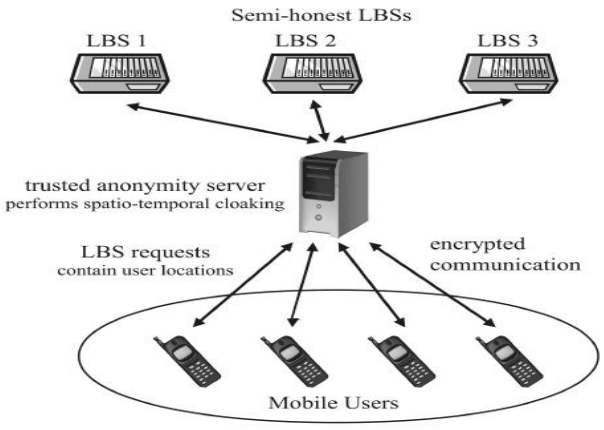

Fig. 1 Architecture of LBS

When a mobile user requests LBS, he will send a user profile to TTP. A user profile is a message defined as follows:

ms $\in S\{uid, nid, (x, y), K, Ct\}$. The payload content in ms is omitted.

As soon as receiving *ms*, TTP divides the whole area into several clusters. The exact location information in *ms* is replaced by MBR of the user's cluster so as to achieve K anonymity. Consequently TTP sends a message *mt* to LBS. Let $\phi(t, s) = [t - s, t + s]$, which extends a numerical value *t* to a range by amount *s*. *mt is* defined as follows: $m_t \in T$: $\{u_{id}, n_{id}, X: \phi(CX, 1/2W_{MBR}), Y: \phi(CY, 1/2\ H_{MBR}), C_{t}\}$.

III. K-ANONYMITY AND LOCATION K-ANONYMITY

Let, RT (A1…An) be a table and QI_{RT} be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in RT $[QI_{RT}]$ appears with at least k occurrences in RT $[QI_{RT}]$. Change data in such a way that for each tuple in the resulting table there are at least (k-1) other tuples with the same values for the quasi-identifier. It is called k-anonymity.

The concept of k-anonymity was originally introduced in the context of relational data privacy. It addresses the question of "how a data holder can release its private data with guarantees that the individual subjects of the data cannot be identified whereas the data remain practically useful". K-anonymity prevents such a privacy breach by ensuring that each individual record can only be released if there are at least k - 1 distinct individuals whose associated records are indistinguishable from the former in terms of their quasi-identifier values.

There are two popular approaches to protect location privacy in the context of LBS usage: Policy-Based and Anonymity-Based approaches. In policy-based approaches, mobile clients specify their location privacy preferences as policies and completely trust that the third party LBS providers adhere to these policies. In the anonymity-based approaches, the LBS providers are assumed to be semi honest instead of completely trusted. We advocate k-anonymity preserving management of location information by developing efficient and scalable system-level facilities for protecting the location privacy through ensuring location k-anonymity.

In the context of LBSs and mobile clients, location k-anonymity refers to the k-anonymity usage of location information. A subject is considered location k-anonymous if and only if the location information sent from a mobile client to an LBS is indistinguishable from the location information of at least k - 1 other mobile clients. The location perturbation is an effective technique for supporting location k-anonymity and dealing with location privacy breaches exemplified by the location inference attack.

IV. PERSONALIZED LOCATION K-ANONYMITY

A. Message Anonymization Basics

In order to capture varying location privacy requirements and ensure different levels of service quality, each mobile specifies its anonymity level (k value), spatial tolerance, and temporal tolerance. The main task of a location server is to transform each message received from mobile clients into a new message that can be safely (k-anonymity) forwarded to the LBS provider. The key idea that underlies the location k-anonymity model is twofold. First, a given degree of location anonymity can be maintained, regardless of population density, by decreasing the location accuracy through enlarging the exposed spatial area such that there are other k -1 mobile clients present in the same spatial area. This approach is called spatial cloaking.

Second, one can achieve location anonymity by delaying the message until k mobile clients have visited the same area located by the message sender. This approach is called temporal cloaking.

It is denoted as the set of messages received from the mobile clients as S. It is formally defined as the messages in the set S as follows [1]:

$$m_{s \in S:} <u_{id},r_{no,\{t,x,y\},K,\{}d_t d_x d_y\},C>$$

The k value of the message specifies the desired minimum anonymity level. A value of k ¼ 1 means that anonymity is not required for the message. A value of k > 1 means that the perturbed message will be assigned a spatiotemporal cloaking box that is indistinguishable from at least k - 1 other perturbed messages, each from a different mobile client. Thus, larger k values imply higher degrees of privacy. One way to determine the appropriate k value is to assess the certainty with which an adversary can link the message with a location/identity association or binding. This certainty is given by 1/k.

The *dt* value of the message represents the temporal tolerance specified by the user. It means that the perturbed message should have a spatio-temporal cloaking box whose projection on the temporal dimension does not contain any point more than dt distance away from t. Similarly, d_x and d_y specify the tolerances with respect to the spatial dimensions. The values of these three parameters are dependent on the requirements of the external LBS and users' preferences with regard to QoS. The set of perturbed messages is denoted as T. The messages in T are defined as follows [1]:

$$m_t \in T:<u_{id},r_{id},\{X:[x_s,x_c],Y:[y_s,y_c],I:[t_s,t_c]\},C>$$

In a perturbed message, X: [xs; xe] denotes the extent of the spatio-temporal cloaking box of the message on the x-axis, with xs and xe denoting the two end points of the interval. The definitions of Y: [ys; ye] and I: [ts; te] are similar, with the y-axis and t-axis replacing the x-axis, respectively. The spatio-temporal cloaking box of a perturbed message is denoted as Bcl(mt)and define it as (mt:X : [xs; xe];mt:Y : [ys; ye]mt:I : [ts; te]). The field C in mt denotes the message content.

B. Message Perturbation: Design and Algorithms

a. Engine overview

The message perturbation engine processes, each incoming message ms from mobile clients in four steps. The first step, called zoom in, involves locating a subset of all messages currently pending in the engine. This subset contains messages that are potentially useful for anonymizing the newly received message ms. The second step, called detection, is responsible for finding the particular group of messages within the set of messages located on the zoom-in step such that this group of messages can be anonymized together with the newly received message ms. If such a group of messages is found, then the perturbation is performed over these messages in the third step, called perturbation, and the perturbed messages are forwarded to the LBS provider. The last step, called expiration, checks for pending messages whose deadlines have passed and thus should be dropped. The deadline of a message is the highest point along the temporal dimension of its spatio-temporal constraint box, and it is bounded by the user-specified temporal tolerance level.

The cloaking algorithms is referred that make their decisions based on this theorem as the CliqueCloak algorithms. The perturbation engine, which is driven by the local-k search as the main component of the detection step, is the base CliqueCloak algorithm.

b. Grouping Messages for Anonymization

A key objective for location anonymization is to develop efficient location cloaking algorithms for providing personalized privacy protection while maintaining the desired QoS. Let us consider this problem in two steps in reverse order [1]: 1) given a set M of messages that can be anonymized together, and 2) for a message ms € S. A set M S of messages are said to be anonymized together if they are assigned the same cloaking box and all the requirements are satisfied for all messages in M.

By considering the second step: Given a message ms € S. Based on the above analysis and observations, one way to tackle this problem is to model the anonymization constraints of all messages in S as a constraint graph defined below and translate the problem into the problem of finding cliques that satisfy certain conditions in the constraint graph.

Definition (Constraint graph): Let G(S, E) be an undirected graph where S is the set of vertices, each representing a message received at the trusted location perturbation engine, and E is the set of edges. There exists an edge e = (m_{si},m_{sj}) € E between two vertices m_{si} and m_{sj} if and only if the following conditions hold: 1) L(m_{si}) € Bcn(m_{sj}), 2) L(m_{sj})€ Bcn(m_{si}), and 3) m_{si} .uid ≠ m_{sj} .uid. It is called as the constraint graph.

Conditions 1, 2, and 3 together state that two messages are connected in the constraint graph if and only if they originate from different mobile clients and their spatiotemporal points are contained in each other's constraint box defined by their tolerance values.

c. Clique Cloak Algorithm

Definition: This algorithm assumes a different k-anonymity requirement for each user. Clique Cloak constructs a graph and cloaks user locations when a set of users forms a clique in the graph.

Taxonomy:

Constraint Area:

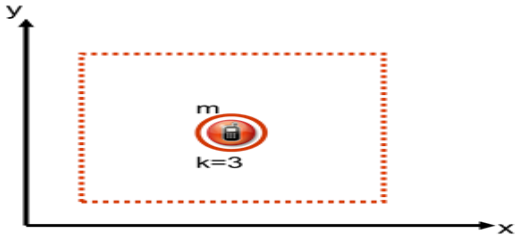

Fig: 2

The fig:2 illustrated that For message m , a constraint area is a spatial-temporal area that contains the sending client location .A client sends his message along with a constraint area to prevent database from sending the client useless information on locations outside the constraint area.

Cloaking Box:

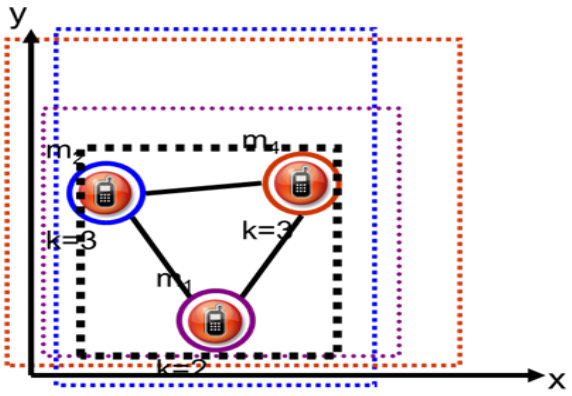

Fig: 3

The fig: 3 illustrated that a spatial and temporal area assigned to a transformed message. A valid cloaking box must comply the following condition:

- The client that sends the message m is located in the cloaking box.
- The number of different client inside the cloaking box must be atleast m.k.
- The cloaking box must be introduced in the message constraint area.

Constraint graph:

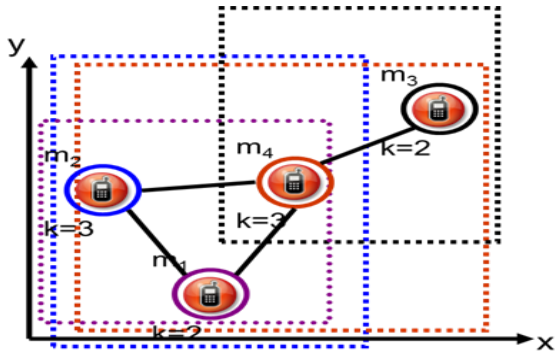

Fig: 4

The fig 4 illustrated that each mobile node is a vertex in the graph and 2 nodes are connected if each of them is inside the other node constraint area.

L-clique:

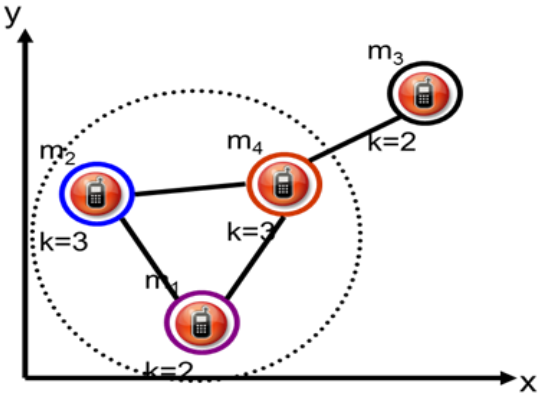

Fig: 5

The fig 5 illustrated that a l-clique in that graph such that $l \geq m_i.k$ for each i is mapped by the algorithm to a spatial cloaking box where all messages in the clique will be transformed using the cloaking box, making each of messages senders indistinguishable from one another.

d. Engine Algorithm

Our location privacy model is the development of an efficient message perturbation engine, which is run by the trusted anonymization server, and performs location anonymization on mobile clients' LBS request messages such as identity removal and spatio-temporal cloaking of location information. We develop a suite of scalable and efficient spatiotemporal location cloaking algorithms, taking into account the trade-off between location privacy and QoS. Our location perturbation engine can continuously process a stream of messages for location k-anonymity and can work with different cloaking algorithms to perturb the location information contained in the messages sent from mobile clients before forwarding any request messages to the LBS provider(s).

The algorithm has 4 phases [1].

- *Phase-1 (Zoom in):* Updating the data structures with new messages from the message queue and integrate the new message into the constraint graph.
- *Phase-2 (Detection):* Applying the local-k search algorithm in order to find a suitable clique in the focused sub graph.
- *Phase 3 (Perturbation):* Generating the k-anonymzed messages to be forwarded to the external LBS providers.
- *Phase 4 (Expiration):* Taking care of the expired messages.

V. CLUSTERING BASED LOCATION PRIVACY

Cluster algorithms are running in TTP, and TTP helps to achieve location K-anonymity according to the required anonymity level of each mobile user. The Cluster Cloak has the following features:

- *K-anonymity*: K in ms means the anonymity level the user desires, which must be met by TTP.
- *High quality QoS:* There is a balance point between QoS and K in the past approach **[**1]. High quality QoS requires the diminution of the cloaked area, which may increase the chance of being found by attackers.
- *High efficiency:* The mobile user's query must be responded by the LBS provider in real time. Even if users move, TTP must finish clusters adjustment quickly.
- *Robustness:* In some cases, though users move, clusters do not need adjusting, which reduces the workload of TTP.

Clustering uses the following definitions:

- *Definition 1*: Cluster Area is a circle whose radius is the distance from the center to the remotest point in the cluster.
- *Definition 2*: Neighbor Clusters are two tangent clusters or two intersecting clusters.
- *Definition 3*: Pneed is the probability of rebuilding a cluster when a mobile user moves.
- *Definition 4*: Nex is the number of extra nodes, without which the cluster can still keep robust.

Cluster Area and Neighbor Clusters are used in the process of cluster merging. Pneed and Nex are utilized to judge if a cluster needs dividing.

The process of building a cluster. After selecting the cluster centre, each point is assigned to the nearest cluster according to the distance from it to the centre. Then new centre will be calculated and each point is assigned to the nearest cluster again. The above process will repeat until the sum distance between every point and cluster centre (CDS) converges to a certain range. The new cluster centre and CDS are calculated as follows.

$$cx = \frac{1}{||C_i||}\sum_{j \in C_i} x_j$$

$$cy = \frac{1}{||C_i||}\sum_{j \in C_i} y_j$$

(x_j, y_j) is the coordinate of a point j in the cluster C_i.

$$CDS = \sum_{j \in C_i} \sqrt{((x_j - cx)^2 + (y_j - cy)^2)}$$

The algorithm used to build a cluster, when the cluster division required.

- *Building Clusters Algorithm:* Building cluster algorithm [2] states the process of building a cluster.
- *Cluster Division Algorithm:* Cluster Division algorithm [2] states the process of cluster division.
- *A user's Partition Algorithm:* A user's Partition algorithm [2] illustrates the process of adjusting a cluster when the cluster size changes i.e. a new user joins the cluster.
- *A cluster's Adjusting Algorithm:* A cluster's Adjusting Algorithm [2] illustrates the process of adjusting a cluster when the cluster size changes i.e. a new user joins the cluster.
- *A user's leaving Algorithm:* A user's leaving Algorithm [2] comes into play when a mobile user leaves a cluster. It finds the cluster in which the user resides and then deletes the user from that particular cluster. Next it goes for adjusting the cluster as per algorithm 1 and algorithm 2. If a cluster does not meet the k-anonymity requirement, then the algorithm goes for cluster merging.
- *Clusters mergence and division Algorithm:* Clusters mergence and division Algorithm [2] illustrates the process of merging two clusters.

VI. COMBINED CLUSTERING SCHEME FOR PROTECTING LOCATION PRIVACY

Cloaking technique group the queries into clusters so as to reduce the energy cost, but lacks of safeguard against the homogeneity attacks and query association attacks. Query protection can well protect those two attacks, but lead to high energy cost in maintaining the diversity of the queries.
The motivation of reducing the complexity while enhancing the performance of the clustering methods in preserving the location and query privacy helps on preservations of location privacy and query privacy of users in continuous location based services in pervasive computing environments. The integrated protection of location privacy and query privacy are merged to form a combined clustering algorithm named as Enhanced Clustering Cloak (*ECC* for short). *ECC* recursively groups all the requests into clusters meeting K−Anonymity and BK−Diversity (i.e. A property that can satisfy l−diversity and m−invariance for the requests) to realize preserving location privacy and query privacy.

A. ENHANCED CLUSTER CLOAK (ECC) ALGORITHM

As illustrated in Algorithm [3], originally all the users are randomly deployed in the region, and then some of the users send their requests to the TTP. The TTP records all of the requests in Qm, then the locations of the requests are recorded in Im. Each request updates its temporal neighbour set when a new user is added.

a. K- means Clustering Algorithm

Algorithm [3] illustrates how K means clustering can be used to divide a cluster into two new clusters. For this it uses the concept of Cluster centre and SSE (Distance of a user from the cluster centre.

b. Hierarchical Clustering Algorithm

Algorithm [3] illustrates the process of Hierarchical clustering for generating cores of clusters. In order to achieve BK- diversity requirement, the clusters should involve all kinds of queries in the universe. Thinking over the essence of the protecting query privacy, the objective is to find a minimum query subset that covers all kinds of queries. This can be achieved by using a hierarchical clustering algorithm. A cluster core is a subset of queries in the clusters, it can be considered as the centre of the clusters in the K-means clustering.

VII. DRAWBACKS

The drawbacks of using a large K-anonymity spatial region, which is an area that encloses the mobile user querying to an LBS server. However, due to the computation overhead of the clique graph, this approach is only able to meet the small K-anonymity requirements of mobile users. When using clustering over LBS'S, that only protects the location privacy of a mobile user, but not protect the query privacy.

VIII. CONCLUSION

Location Privacy is a particular type of context privacy. It is defined as the ability to prevent other unauthorized parties from learning one's current or past location. The personalized location k-anonymity is discussed. Several variations of the spatio-temporal cloaking algorithms, collectively called Clique Cloak algorithm are the core algorithms for perturbation engine. A location privacy preserving scheme for pervasive computing environment named Cluster Cloak satisfy the privacy and QoS requirement of the users. Cluster Cloak is adopted by TTP, and clusters can be adjusted in real time when users move from one domain to another domain. A combined clustering algorithm (*ECC*) for preserving the location privacy and query privacy. *ECC* runs on the TTP, it applies an iterative K-means clustering method to group the user requests into clusters for providing location safety while utilizing a hierarchical clustering method for preserving the query privacy. *ECC* provides the mobile users with their desired anonymity levels and spatial tolerances.

REFERENCES

1. B. Gedik and L. Liu, "Protecting location privacy with personalized k-anonymity: Architecture and algorithms," Mobile Computing, IEEE Transactions on, vol. 7, no. 1, pp. 1–18, 2008.
2. L. Yao, C. Lin, X. Kong, F. Xia, and G. Wu, "A clustering based location privacy protection scheme for pervasive computing", in Proceedings of the 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing. IEEE Computer Society, 2010, pp. 719–726.
3. Chi Lin, Guowei Wu, Lin Yao, Zuosong Liu"A Combined Clustering Scheme for Protecting Location Privacy and Query Privacy in Pervasive Environments", 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications
4. M. Gruteser and D. Grunwald, "Anonymous Usage of Location-Based Services through Spatial and Temporal Cloaking", Proc. ACM Int'l Conf. Mobile Systems, Applications, and Services (MobiSys '03), 2003.
5. L. Liu, "From Data Privacy to Location Privacy", VLDB '07: Proceedings of the 33rd international conference on Very large data bases, ACM Press, Sep. 2007, pp. 1429-1430.

AUTHORS PROFILE

Ipsa Das received his B.Tech degree in the year 2012 from BPUT, Rourkela and currently pursuing M.Tech in Computer Science Engineering from KIIT University, Bhubaneshwar. Her research area includes Data Mining, Clustering and Database Systems

Md. Imran Alam received his B.E degree in the year 2011 from GTEC, Vellore affiliated to Anna University, Chennai, and currently pursuing M.Tech in Computer Science Engineering from KIIT University, Bhubaneswar. His research area includes Distributed Database systems, Task Scheduling, Data Mining, Cloud Computing and Big Data.

Jayanti Dansana is presently working as an Assistant Professor in the School of Computer Engineering, Kalinga Institute of Industrial Technology, and Bhubaneswar. Her research interest areas include Data Mining, Clustering, and K- means.