World Wide Web (2020) 23:175–206
https://doi.org/10.1007/s11280-019-00716-7

Preserving location privacy using three layer RDV masking in geocoded published discrete point data

Ruchika Gupta[1] · **Udai Pratap Rao**[2]

Received: 8 February 2018 / Revised: 24 July 2019 / Accepted: 30 July 2019 /
Published online: 13 August 2019

Abstract
The prevalent usage of Location Based Services; where getting any informational service is solely based on the user's current location, have raised an extreme concern over location privacy of the user. The privacy concern becomes paramount when the location tagged data publication like government health care data, district crime record data and the like, are reverse engineered by an adversary to pinpoint the real user against the location given in the specific tuple of the record. Address information is typically considered as a confidential element of the published record and any linkages of this piece of information with publicly available quasi identifier is enough to reveal a lot about a user (which is not apparent otherwise) or hamper the social reputation of the user considering the extreme case. Various geographical masking techniques have been presented and discussed at length in the literature, however, no scheme is able to dispense privacy providing absolute usefulness of the published data. This work is a research attempt to recognize the current state-of-the-art in geographical masking, supportive analysis of the existing masking technique, and come up with a robust solution that serves the purpose of location privacy without making published data worthless. The suggested solution is well suited for geocoded, static, *discrete point* published data.

Keywords Geocoded published data · Location privacy · Geomasking · User privacy

1 Introduction

The intense development of location detection empowered devices and escalating availability of wireless interconnections almost everywhere results in emerging location based applications. In Location Based Services (LBS) we incline to use positioning technology to register mobile location movement. There are quite a lot of abstract approaches and real implementations of systems to resolve the place of a cell phone. The most outstanding example of such a positioning system is the GPS [22]. Although LBS offers major openings for a

✉ Ruchika Gupta
rgupt009@gmail.com

[1] Department of Computer Science and Engineering, Chandigarh University, Punjab, 140413, India

[2] Computer Engineering Department, National Institute of Technology, Surat, Gujarat, 395007, India

large variety of markets and remarkable convenience to the end user, it also presents subtle privacy attacks at the same time. Privacy of the system is threatened due to the requirement of the current location of the user in order to provide the related services.

The basic aim in the process of location privacy research is not only to save the privacy/ security of the user but also to publish and share important results with other communities like researchers, analyst, and the general public. An access to the shared result enables the analyst to link the results with the auxiliary data that helps to take a decisive safety measure moves after facilitating enhanced knowledge production. Geoprivacy is considered as the privilege to decide how and when one's personal location data is shared with other parties [1, 10, 25].

1.1 Geocoding

Geocoding is the technique used to convert the addresses into their corresponding latitude and longitude combinations. Using geocoding, an entire address with the information of residence number, street code, city name, and the state information gets converted into a precise dot over a map (including other cartographic material) to represent a given user. Hence, the latitude and longitude field in the data records and corresponding points over map required to be treated as highly confidential information that need to keep private, failing to do so may end up revealing much about a user which he would not disclose otherwise. Such accidental release of data may even hamper a social reputation of the user in an extreme case, for instance, Alice would never want to reveal that she suffers from a fatal disease like AIDS or even her single visit to an AIDS clinic she would like to keep secret. Public discernment is another very important aspect to contrivance different geo-masking techniques in the data releases. People may object if their entities are replaced by a sharp dot in the map and may consider it as a pure privacy betrayal. Therefore, it is the foremost responsibility to bring a good-faith effort that is able to protect user's identity.

Sophisticated tools are available that perform the reverse geocoding to produce an approximate address location based on the given latitude, longitude information in the released record. A data record with actual location contents (notice that real names of the patients are either removed or masked) can be reverse engineered with the help of such sophisticated tools to identify the location and the user inference can easily be made thereafter.

1.2 Purpose of geomasking

Geomasking (or geographic masking) is the method of altering the actual location coordinates of the area with a specific purpose to limit the risk of re-identification of the entities after data release. This term was first coined by [2] as an extension of masking techniques used for non-spatial microdata [6, 11]. The underlying purpose of the work done in the field of geomasking followed by location analytics is to protect the real world identity of a particular entity, consider health subject for instance where identity of a patient can be revealed comfortably from the given patient's address. Therefore, address attribute ought to be treated as a confidential information and supposed to be removed before the actual data publication releases take place. Removal of the addresses from the released data leads to weaker data utility from a research perspective. Various analysis can be helpful to discover the cause of the disease pertaining to the certain demographics to the intent that safety measures can be planned accordingly and location masking is one of such techniques that helps to hide the real location of the user. Hence, the complete address data removal is replaced by masking the location data in which the actual geo-coded data is perturbed/ masked in order to keep the released data meaningful and research oriented.

Geomasking is generally performed for one of the following two types of data releases:

1. For tabular data release that involves subsequent analysis, and
2. For public presentation of a map

Here, in the first case the basic aim is to preserve the spatial relationships of the attributes so that the analysis does not get affected by the absence of it while to provide an visual pattern analogues to the given data after geomasking is the main aim in later case.

1.3 Benefits and risk involved with data sharing

Mapping of the georeferenced user data with the spatial analysis can help in identifying the profound geographical patterns or lead to a significant learning for managing particular social issues in a specific territory.

There are few considerations that must be balanced before releasing the location information of a user.

a. The requirement of the user's confidentiality protection,
b. Original patterns should be preserved, and
c. The usefulness of the sharing data released

The protection of the user's confidentiality is the primary condition, also a fundamental right of an individual's privacy. Data sharing should be done in a manner that does not dissolve the usefulness of the information and the original pattern remains intact for the benefit of the research community which in turn results into the real benefit for the public at large. The above mentioned considerations are difficult to achieve as, on the one hand, to attain high confidentially the masked changes should be maximized and, on the other preserving the original patterns is achievable by minimizing these changes. The main objective is to maintain a balance between the two so that the risk of re-identification becomes as low possible without changing properties of the real data records. In order to achieve these goals we propose a masking solution called 'Three Layer RDV Masking' that makes the record location attribute de-identification difficult. Here, the RDV refers to the three layers namely, **R**egion, **D**elaunay, and **V**ornoi we propose in the solution.

The rest of the paper is organized as follows: Section 2 highlights the related work in the field of location privacy. Sections 3 discusses the methodologies used while proposing the solution. Section 4 exhibits the proposed three layer RDV masking solution and proposed algorithm. Section 5 presents performance metrics of RDV Masking solution. Finally, Section 6 concludes the paper.

2 Related work

Most of the location privacy work has been carried out for location based services setup where a user sends her location in order to get the related location service [18]. Getting an optimum quality of service without compromising the privacy is a big concern and various techniques are proposed to preserve the location privacy of the user in the mentioned scenario [17, 19, 20]. In the course of the last couple of years, there has been a surge of enthusiasm for geoprivacy among geographic groups [14, 40]. The geoprivacy right is not always taken as location protection right as no direct location data is sent to the service provider, however, location information is tagged with some other data in the form of a supported attributes. Authors in [24] suggest that the new spatial techniques, absence of

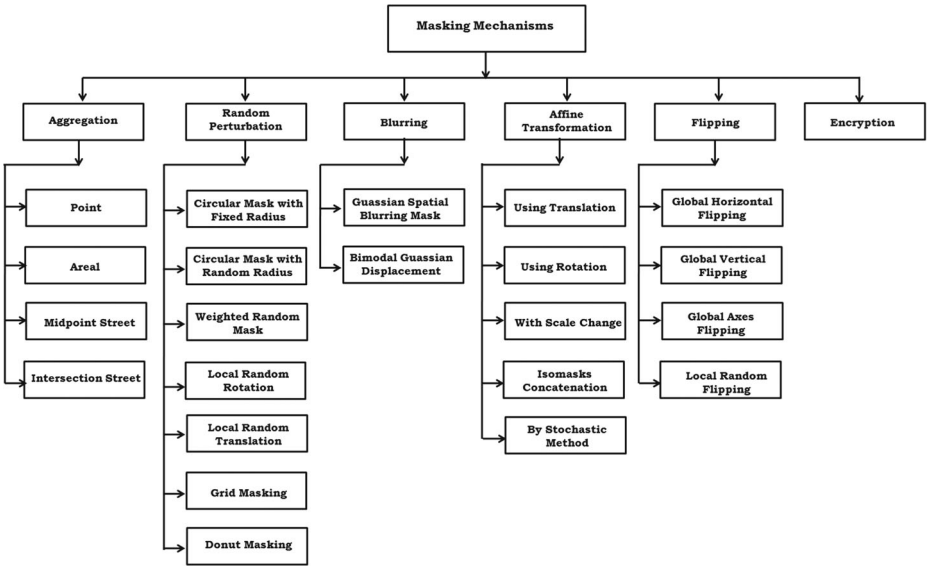

Figure 1 Classification of existing geomasking techniques

well designed specific laws, and laxity in the published records can be the major reasons responsible for location disclosure.

In literature there are many techniques/ schemes/ algorithms proposed by numerous practitioners and researchers and these techniques can be broadly summarized in the form of a tree like grouping classification model shown in the Figure 1.

Aggregation is one of the very basic methods for geomasking. Aggregating data points to the specified territories before releasing the actual records is the basic methodology of aggregation masking technique. Here, the spatial resolution is reduced and the produced patterns are no closer to the original data patterns, therefore, the usefulness of the published records hamper. The high spatial resolution helps to detect the underlying patterns, such as crime rate, disease risk, epidemic outreach, and the like [25]. Following are the different aggregation techniques discussed in the literature at length. In point aggregation technique, a new point replaces the location of several other points [2] and provides a variable masking degree shown in the Figure 2.

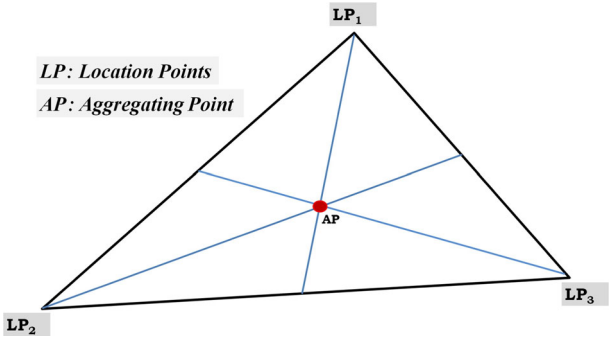

Figure 2 An instance of point aggregation

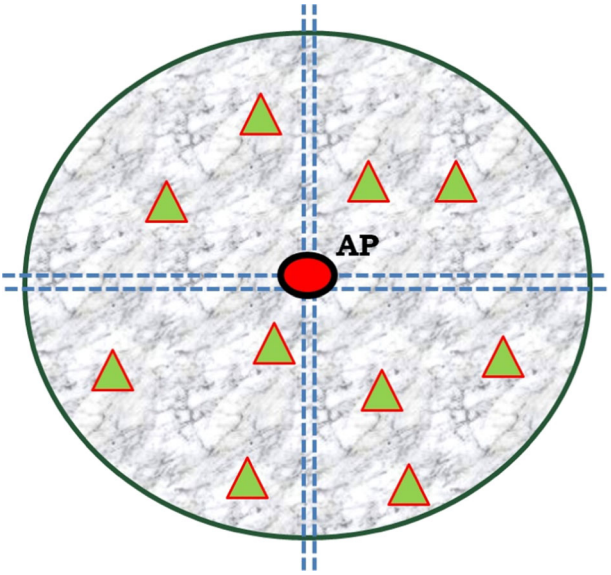

Figure 3 An instance of areal aggregation

In areal aggregation, a new symbolic area portrays the location replacement of the several location points possess variable masking degree [2] shown in the Figure 3. The method of aggregation at the midpoint of the street involves the aggregation at an aggregating point (AP) that is represented by the center point of the street [27, 28]. The Figure 4 shows an instance of the mechanism.

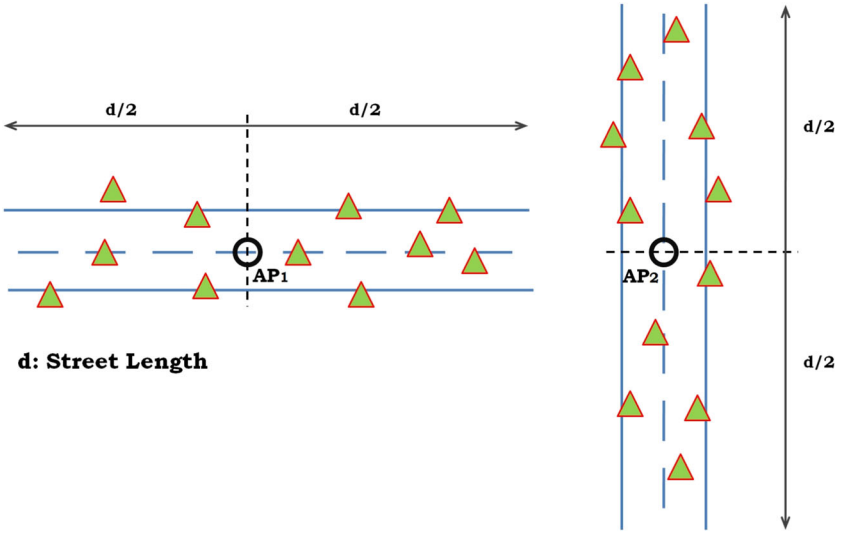

Figure 4 An instance of street midpoint aggregation

In aggregation at the intersection of a street, AP for several location points lies at the point where streets crossed to each other [27, 28] as shown in the Figure 5. This technique was first applied over the homicide data of the region and used for disease cases clustering.

Random Perturbation is a rudimentary method used to control data disclosure that works on the principle of adding a random value (or noise) to the original data value and the resultant value is substituted as the original value in order to reduce the re-identification attack. Following are a few types of random perturbations. In circular masking with fixed radius method, the original location point is replaced by another location point produced after adding a fixed displacement in a circular fashion. Though the direction of displacement is not fixed and can be random within the specified radius [25] as shown in the Figure 6. Circular masking with random radius method is similar as circular masking with fixed radius, the only difference is that it allows the radius displacement value to be random in random direction which lies between $l_o \leq d_i \leq d$, where l_o is the original location point, d_i is the random displacement chosen for the case, and d is the radius upper bound [25]. Since all possible locations within the circle are equally likely, therefore, masked locations are more likely to be placed at a larger displacement distance compared to the smaller distances as shown in the Figure 7. Another method called the weighted random mask works same as circular masking with random radius but here the value of the radius depends over the population density of the masked area [25]. It can be observed that all of these methods translate the location point far from the original point by a displacement inversely proportional to the population density of that area. Let $\mathcal{P}$ be the populace density of area $\mathcal{A}$ and let d describes the displacement, then following expression holds the condition for the value of displacement threshold and given as,

$$\mathrm{d} \propto \frac{1}{\mathcal{P}}$$

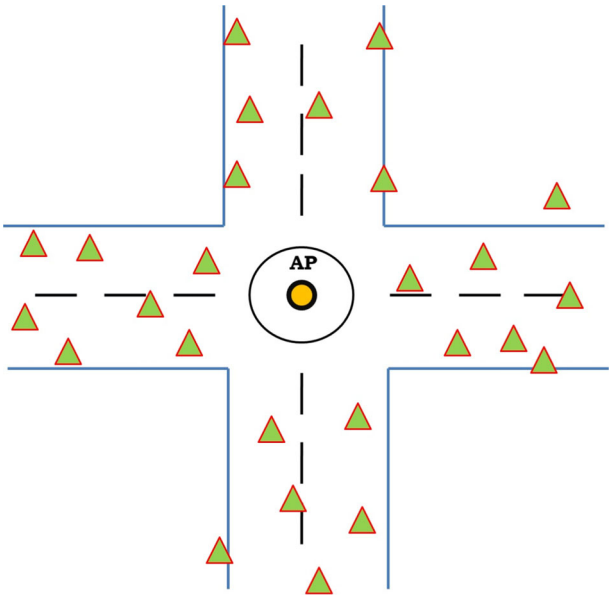

Figure 5 An instance of street intersection aggregation

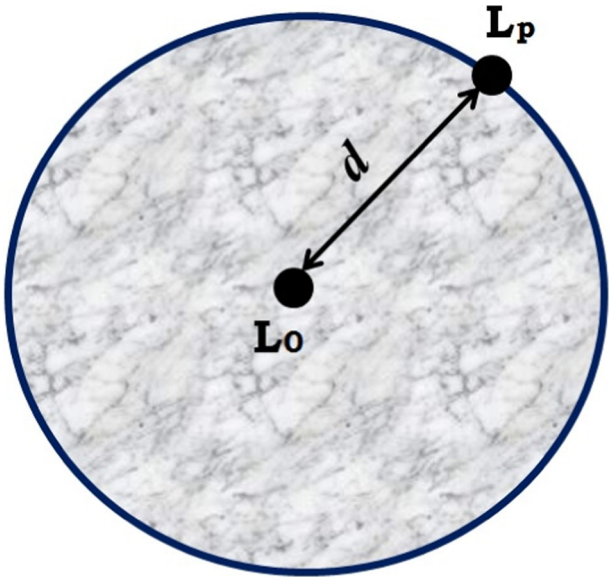

Figure 6 Illustration of circular masking with fixed radius

Local random rotation technique rotates the center of each grid cell of original location point. The masking degree of this technique is fixed for all perturbations [28] (refer Figure 8).

In local random translation, each location point is translated by a random translation factor δ within a grid cell [27] (refer Figure 9).

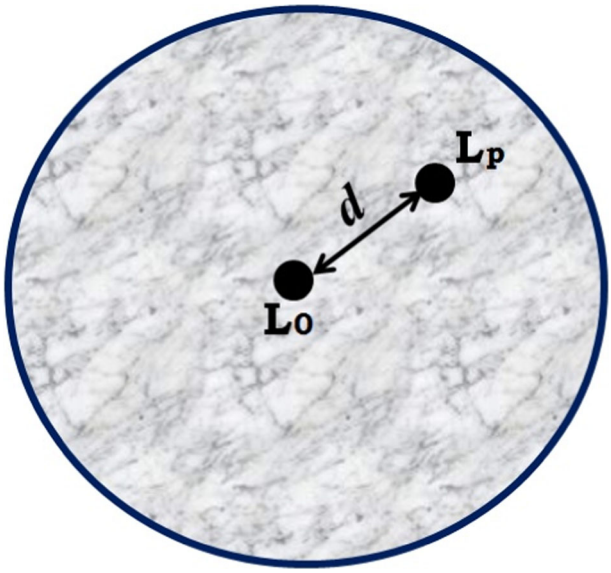

Figure 7 Illustration of circular masking with random radius

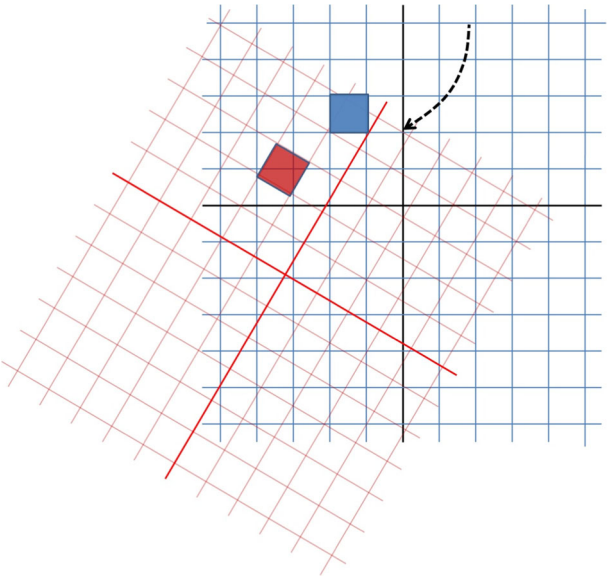

Figure 8 Illustration of local random rotation

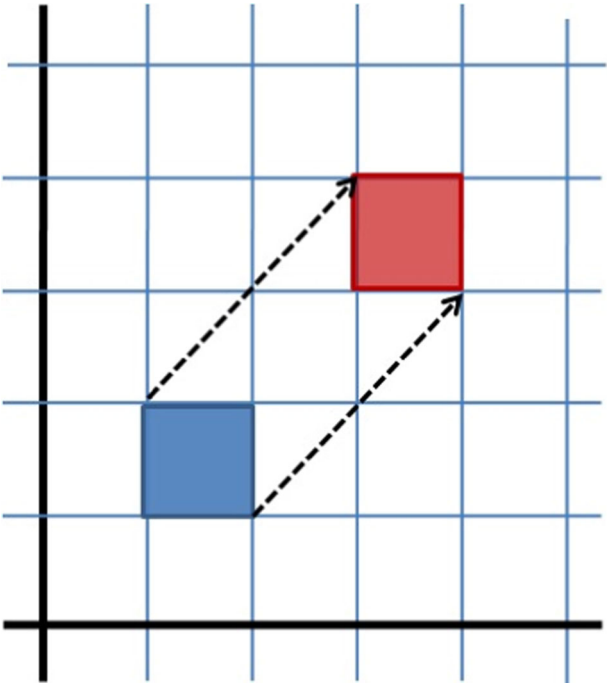

Figure 9 Illustration of local random translation

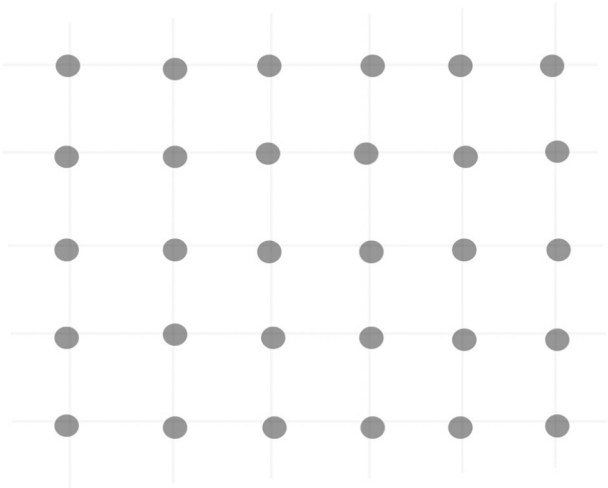

Figure 10 Illustration of grid masking

Grid masking disparate every original location data point to uniform grid cells [7, 28] as shown in the Figure 10.

In donut masking each location point is displaced by a random displacement value bound within a minimum and maximum threshold [21] shown in the Figure 11.

Blurring is the technique to make the data points less distinct to reduce re-identification of the user location and hence the user's real world identity. In Guassian spatial blurring

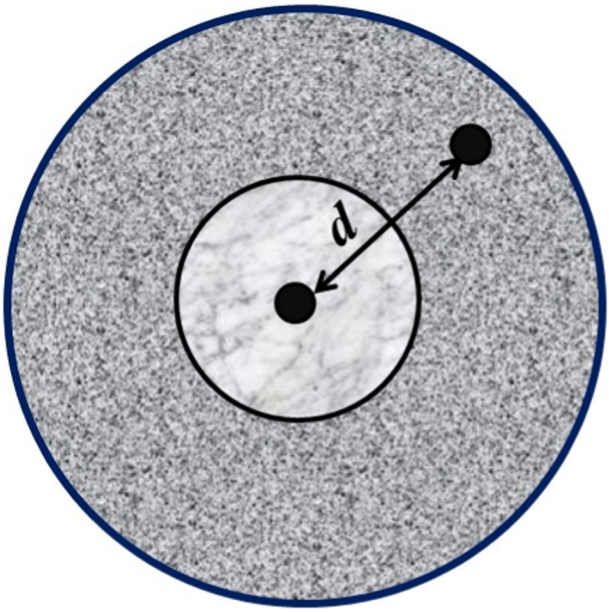

Figure 11 Illustration of donut masking

mask, the displacement direction of location point is irregular, however Gaussian distribution is used for the separation computation. The scattering of the dissemination can be differed on the basis of important parameters like neighborhood populace density [4] shown in the Figure 12.

Bimodal Guassian displacement technique is a variation of Guassian blurring, utilizing a bimodal Gaussian distribution for random distance computation [4] shown in the Figure 13.

The methods under affine transformation technique suggest the use of basic geometric transformations to displace the location point [2]. Degree of masking in such methods are always constant and does not support the presence of other parameters like population density of the region. Translation method involves the displacement of the location point by a constant translation factor. Rotation method rotates the location point by a fixed constant angle about the origin. In scale change method, the point is displaced by a fixed scaling factor. Isomasks concatenation is a hybrid strategy, suggests the displacement of location points by the combination of translation/ rotation/ and scaling with a fixed transformation factor. Stochastic method allows the predefined random range of the transformation factors for translation/ rotation/ scaling of location points.

Flipping method is based on the tossing of axis locally and globally. Following are a few types of flipping masking method [27, 28]. a. Global Horizontal Flipping method flips the horizontal central axis of the map to mask the location point. b. Global Vertical Flipping method flips the vertical central axis of the map to mask the location point. c. Global Axes Flipping method flips both the axes i.e. horizontal as well as vertical central axes of the map to mask the location point. d. Local Random Flipping method flips horizontal, vertical, or both the axes of each grid cell of the map randomly to mask the location point.

Transmission of data can also be made secure by encrypting (using public key or private Key cryptosystem) it along with the enforced laws, well defined rules and standard regulations [39]. The Constitution of India does not patently grant the fundamental right to

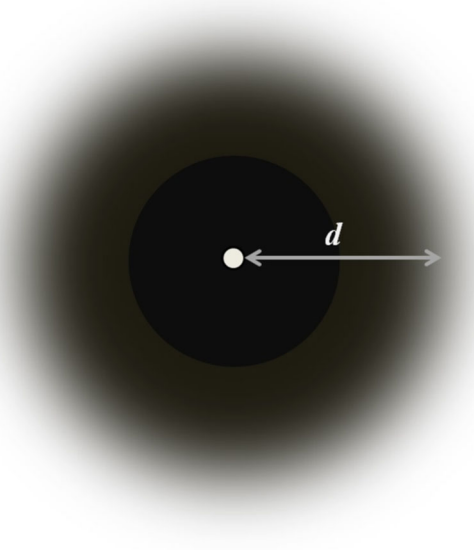

Figure 12 Illustration of Guassian spatial blurring mask

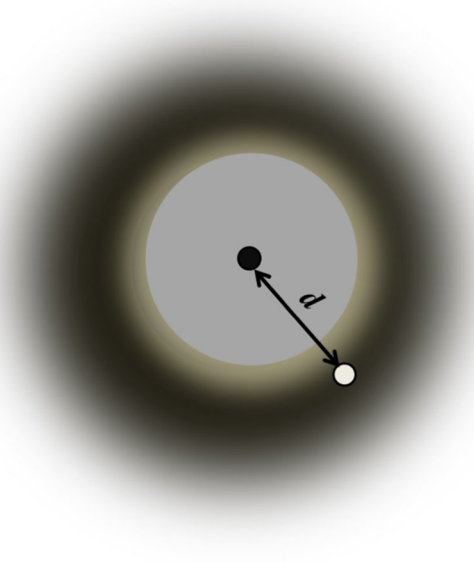

Figure 13 Illustration of Bimodal Guassian displacement

privacy. However, the courts have read the right to privacy into the other existing fundamental rights; i.e., freedom of speech and expression under Article 19(1)(a) and right to life and personal liberty under Article 21 of the Constitution of India. India presently does not have any express legislation governing data protection or privacy. However, the relevant laws in India dealing with the data protection are the Information Technology Act, 2000 and the (Indian) Contract Act, 1872 [12]. A codified law on the subject of data protection is likely to be introduced in India in the near future [18]. It becomes the responsibility of publishers and researchers who releases the data record to ensure the location privacy of the individual, if legal practices are inadequate/ fails.

The following are a few observations that relate the methodologies with the problem we have identified.

1. Most of the solutions in geomasking focus only on the population density as a weighted parameter of the region used for masking.
2. Majority of the suggested schemes are unable to intact the original data pattern and the masked data patterns has no resemblance to the original data pattern. The published data record with masked data which has no resemblance or very vague resemblance to the original data pattern is of no use to take right decisions. Such type of data publication is considered useless from analysis stand point and provides no help to the research communities.
3. The masking degree can be steady or variable. It is said to be constant degree if the measure of the instability range is equivalent for all points. On the contrary, masking is said to be variable if extent of the vulnerability region that fluctuates relies upon a particular factor. Mostly this factor is the underlying population density of the region.

None of the techniques discussed in the literature has put forth an absolute geomasking privacy solution, yet several of them provided efficient mechanisms to provide user privacy

in different scenarios [33, 38] . In recent years many of the static data releases techniques are also proposed [29, 41, 42]. Some solutions are well suitable to preserve original data point pattern while others are good at providing data point masking. The major benefit of our proposed RDV solution lies in the fact that apart from providing efficacious point masking, it also preserves the spatial pattern of the original data points considering various iterative indicators other than only population density.

3 Methodology used

This section presents the methodology used while proposing the solution algorithm.

3.1 Triangulation

In geometry and trigonometry, triangulation can be described as the process of determining the location of a point by shaping triangles to it from given known points [31]. There are various methods of triangulation that exist in the literature, out of which the Delaunay triangulation has certain advantages that can be useful to exploit the geoprivacy protection of the individual in the published data record.

3.2 Delaunay triangulation

A triangular network with irregularity is a decent way to visualize a real world surface morphology. Triangulation of point is computed to construct a vector based model for analysis. Non-overlapping sequence triangles modeled as a huge connection of networks and every spatial region can be triangulated over a given set of points.

Let P be the set of point given in the plane, the Delaunay triangulation DT(P) is formed such that no point $P_i \in P$ lies inside the circumcircle of any triangle [36]. Or in other words, Delaunay triangulation is a proximal strategy that fulfills the necessity that a circle drawn through the three points of a triangle contains no other point. A triangulation of P is legal if and only if it is a Delaunay triangulation. The Figure 14 presents a random instance of Delaunay triangulation method. The geometric center of data points [13], shortest of possible lines between two data points, a line segment of imaginary boundary, and a line segment on the boundary convex hull are determined to find the proximity of a point. Boundary of a Delaunay triangulation is a convex hull produced taking set P as input where shortest Delaunay edge connects the closest points pair.

3.3 A few important properties

Let p be the total points and *dim* be the total dimensions, following are some useful properties of Delaunay triangulation:

1. Union of all triangles produces the convex hull of the points.
2. Delaunay triangulation DT(P) contains $O(p^{dim/2})$ triangles.
3. In the plane where $dim = 2$, if there are b vertices on the convex hull then any triangulation of the points has at most $2p - 2 - b$ triangles and one exterior face.
4. In the plane, each vertex has on an average six surrounding triangles.

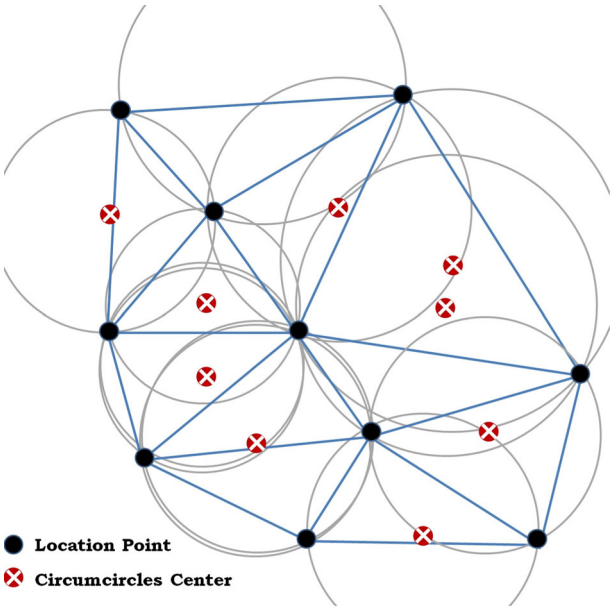

Figure 14 An instance of Delaunay triangulation

5. Delaunay triangulation maximizes the minimum angle in the plane. Compared to any other triangulation of the points, the smallest angle in the Delaunay triangulation is at least as large as the smallest angle in any other.
6. A circle circumscribing any Delaunay triangle does not contain any other input points in its interior.
7. If a circle passing through two of the input points doesn't contain any other of them in its interior then the segment connecting the two points is an edge of a Delaunay triangulation of the given points.
8. The closest neighbor b to any point p_i is on an edge bp in the Delaunay triangulation since the nearest neighbor graph is a subgraph of the Delaunay triangulation.
9. The shortest path between two vertices, along Delaunay edges, is known to be no longer than $\frac{4\pi}{3\sqrt{3}} \approx 2.418$ times the Euclidean distance between them.

3.4 Existing algorithms

There are successful implementations that exist in the literature to achieve Delaunay Triangulation. Following are some of its most popular implementations:

a. **Flip Algorithm** One of the edges can be flipped if the triangle is not a Delaunay triangle. This property leads to a flipping algorithm given by [23] to construct a Delaunay triangle and involves $\Omega(p^2)$ edge flips.
b. **Incremental Algorithm** This algorithm uses a straightforward way for Delaunay triangle computation which repeatedly adds one point at a time. The overall computation time of the algorithm is $O(p^2)$. Points addition can also be performed in a random order

[16]. When the method involves more dimensions the run time is exponential to the dimensions irrespective to the size of resultant Delaunay triangulation [8, 12].

c. **Divide and Conquer Algorithm** Here the line is drawn recursively to split the points into two different sets and then DT(P) is computed for each set. After this computation merging of sets take place to combine the results. Splitting takes $O(\log p)$ and merging takes $O(p)$, hence overall complexity can be given as $O(p \log p)$ [26]. A conscious choice of points can reduce this time to as small as $O(p \log \log p)$ and considered as fastest among all strategies [5, 35].
d. **Sweephull Algorithm** This is a hybrid strategy, uses a gradually sweeping of sorted radius with a flipping algorithm till all triangles satisfy Delaunay [34]. The technique is useful only for 2-D Delaunay triangle computation and does not support higher dimensions.

3.5 Voronoi polygons

Voronoi polygons define those regions where the boundaries are equidistant between the surrounding points. This can be viewed as the areas where inside of the polygons is closer to the corresponding point than to any other point [3, 37]. If the centers of the circumcircles present in Delaunay triangle are connected together, they result into the Voronoi diagram consisting Voronoi polygons as shown in the Figure 15.

Voronoi diagram is a useful geometrical method supports the building of the point location data structure. This data structure is extremely helpful for answering *nearest neighbors queries*, the basic feature of location based services where the queries for instance 'Find the nearest ATM', or 'Find the nearest hospital' are sent by the user frequently.

3.6 Existing algorithms

Following are existing Voronoi diagram computation algorithms:

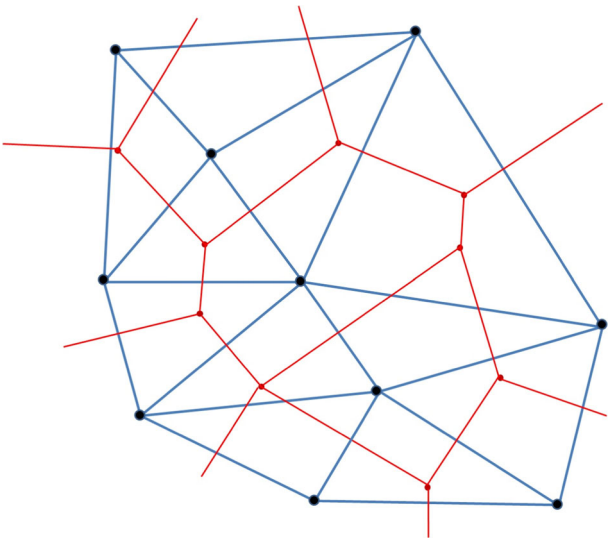

Figure 15 Corresponding Voronoi polygon formation

a. **Fortune's Algorithm** This algorithm is given by [15] for producing Voronoi diagram from the given set of point V. Set V contains centers of circumcircles as the points. The complexity of the algorithm is describes as $O(\mathrm{v}\log \mathrm{v})$.
b. **Lloyd's Algorithm** This algorithm starts with Delaunay triangles as input and produces Voronoi diagram [30]. Authors in [9] presents a generalization of algorithm given by [30] in which Voronoi tessellation is computed.
c. **Bowyer-Watson Algorithm** This method generates Delaunay triangles for any dimension of input and also produces Voronoi diagram from $O(\mathrm{v}\log \mathrm{v})$ to $O(\mathrm{v}^2)$ time [32].

4 Proposed three layer iterative RDV masking

The proposed three layer RDV Masking model is built upon the concept of rudimentary triangulation and Voronoi generation. The suggested scheme is proposed to be executed at research analysis level before releasing the final data to the public. The proposed masking scheme appears to be centralized as the masking is done after inputting the nodes (that are to be masked) at one point. None of the techniques discussed in the literature has put forth an absolute geomasking privacy solution, yet several of them provided efficient solutions for protecting the user privacy. Some of the solutions are suitable to preserve original data point pattern while others are good at providing the data point masking. The major benefits of our proposed RDV solution lies in the fact that apart from providing efficacious point masking, it also preserves the spatial pattern of the original data points considering various iterative indicators other than just population density.

4.1 Layers description

RDV masking works in three layers as described below:

Layer 1: Layer R This is the first phase of the algorithm in which points get spatially distributed over the region. Let N be the given set of points that need to be masked for a given region R. In this phase N is distributed over R and the original pattern of points is stored for difference analysis.

Layer 2: Layer D Second phase of the algorithm (called as Delaunay Layer phase) designates to compute delaunay triangulation of point set N. Divide and Conquer mechanism is used to compute delaunay triangulation that takes total of $O(\mathrm{N}\log \mathrm{N})$ computation time, where N is the given point set.

Layer 3: Layer V Voronoi layer is the third masking layer where every point gets separated to the closest segment along the edges of the corresponding Voronoi polygon. The profound benefit of this layer is plausible in those segments where the original point density is higher, the points are moved a shorter distance in their corresponding masked pattern that appears very much like the original data point pattern. Fortune's Algorithm is used to compute Layer V masking and takes total of $O(\mathrm{N}\log \mathrm{N})$ computation time, where N is the given point set.

4.2 Iterative indicators

Mostly, all geomasking mechanisms utilize region's populace thickness as the main standard for ascertaining the displacements between the original data point locations and the

masked data point locations. Following are few index factors we have identified, which can be considered as iterative indicators (or $\mathcal{I}$) to decide the degree of revealed data points.

a. ***Population Density***
   Risk of re-identification of an individual increases when the region is sparsely populated. Population density is the standard attribute to measure the degree of displacement in the masking mechanisms. Reverse engineering is highly difficult in densely populated area compared to the area with sparse populace.

b. ***Type of Investigation***
   This iterative indicator presents the type of investigation carried out over the released records. Conventional searches involve the basic statistics analysis, therefore, legitimate degree of masking is enough to perform such analysis. On the other hand, advanced progressive analysis requires the least masked data in order to produce more accurate analysis results.

c. ***Statistics Responsiveness***
   This indicator accounts the seriousness of the data, the situation where safety measures at an absolute time line of the event outweigh the stringent privacy need of the user. For example, epidemic maps and fatal diseases like HIV maps.

d. ***Degree of Quasi Identifiers Existence***
   Quasi identifiers are those publicly available attributes that can increase the risk of user re-identification if linked together. Any stand alone quasi identifier does not create any problem to the published record, however, the combination of already existing quasi identifier of the record can be highly detrimental and also make record's re-identification easier.

e. ***Target User***
   A straightforward indicator, that involves the party to whom the published data is catered for analysis. Masking degree differs for the different target users. Users can be general public, a well formed research task force for a specific cause, academic researchers, or a pharmaceutical medicine launcher team.

4.3 Privacy measurement

The original dataset faces a privacy distortion once the suggested methodology is applied to the given discrete points. Here, we theoretically formalize the level of user location point privacy distortion with some definition related to the addressed problem and the proposed solution.

Definition 1 *Original to displaced Discrete Point Dataset*
The data points are displaced from the original position by adding certain displacement value based on the granularity threshold of the selected iterative indicator $\mathcal{I}$. The dataset with the amended values is considered to be a randomized procedure depicting a sequence of location data points arranged in the fashion of selected $\mathcal{I}$. Therefore, the modified data point set P' can be defined as the addition of original data point set P' and the random displacement value within the range of iterative indicator $\mathcal{I}$ which can be given as,

$$\mathrm{P}' = \mathrm{P} + R \tag{1}$$

where R represents the added displacement to the original data point set P.

Definition 2 *Entropy based Privacy Measurement* The user identity and other sensitive information disclosure can be performed by the linking process where the adversary attempts to link the attributes of the publicly available data (mostly from the data present at different sources) with the published records. In such scenarios, the concept of entropy is preferably used to measure the level of privacy. Here, the data is viewed as the sequence of samples which are independent to each other and categorized over some attributes.

Given an estimate and displaced data point set $\hat{\mathrm{P}}$ and P', the entropy based measurement of the privacy $\mathcal{L}(\cdot)$ can be defined as,

$$\mathcal{L}(\mathrm{P}', \hat{\mathrm{P}}) = -\sum_{\mathrm{P}', \hat{\mathrm{P}}} Dist(\mathrm{P}'(k), \hat{\mathrm{P}}(k)) \cdot \ln p[Dist(\mathrm{P}'(k), \hat{\mathrm{P}}(k))] \tag{2}$$

Here, we consider $Dist(\mathrm{P}'(k), \hat{\mathrm{P}}(k))$ as the function defining normalized distance that produces distance in the interval of [0, $\mathcal{I}$], where value 0 depicts *'no privacy'* while $\mathcal{I}$ represents full protection of privacy. We use Euclidean function to evaluate the distance, however different other distance functions can be employed.

In the data masking, entropy depicts the uncertainty while distinguishes a user's real location among other masking points of the region. In our proposed solution the data points are masked with three different layers namely; Layer R, D, and V in which Layer R is responsible to distribute the data points (that need to be masked) over the given region R while rest of the two layers perform the veritable masking and displace the data points based on the selected iterative indicators in order to mask them. Entropy based privacy measurement is relevant only when the data points are moved from their actual positions, hence the two layered privacy is introduced only in the masking Layer D and Layer V, respectively, while the need of privacy is not exercised during Layer R. Utility value of the moved data points is relatively high if the yielded masked data points are similar to the real data points, henceforth the data points utility is maximum when the iterative indicator $\mathcal{I}$ value is chosen to be as low as 0. On the other hand, achieved privacy is maximum when the moved data point set is entirely independent and shares no similarities with the original data point set given.

4.4 The algorithm

Considerations and Assumptions

a. *The algorithm is executed at a highly trusted third party (such as a government organization) before the actual data release.*
b. *Published data is used for analysis in order to take some decisive measures/ actions.*
c. *Static data of the users are collected from an authorized source for an experimental purpose.*
d. *Any entity having access to the released data record can act as an adversary.*
e. *Region* R *is browsed and obtained from OpenStreetMaps application.*
f. *The value of* $\mathcal{I}$ *is specified by the data releasing authority on the basis of severity of indicators.*

The algorithm also takes into account the scenario when the released data is in the form of mass notification and end users are mainly interested in the notifications only and less concerned about the spatial detailing of the published record.

Algorithm 1 Iterative three layer RDV masking.

Function: Masking in three layers using RDV Mask
Let data record has N *geotagged data points that need to be masked*
Let R *be the region of* N *and let converted set of points is given by* $N' = \emptyset$
Let $\mathcal{I}$ *be the iterative indicator*
//Layer 1: Layer R
return N′
for point n ∈ N′ **do**
Distribute over the specified region R
end for
Let DT *be the Delaunay set of points and* VP *be the set of Voronoi points*
Initially, DT = VP = ∅, $i = 1$
if $\mathcal{I} == \langle MAXIMUM\_THRESHOLD \rangle$ **then**
CALL MAX_I_LAYER_Computation(N′);
break;
end if
else
while $i \leq \mathcal{I}$ **do**
//Layer 2: Layer D
CALL Layer_D_Function(N′);
//Layer 3: Layer V
CALL Layer_V_Function(DT);
$i+ = 1$
end while
return DT **return** VP
end

Algorithm 2 MAX_I_LAYER_Computation.

Function: Mean Computation
Let X and Y be the point coordinates of points in the set N′
for $x \in X_p(N')$ and $y \in Y_p(N')$ **do**
i=1, $Temp_x = Temp_y = 0$
while $i <= |N'|$ **do**
$Temp_x = Temp_x + x_i$;
$Temp_y = Temp_y + y_i$;
i++;
end while
$X_m = Temp_x/|N'|$, $Y_m = Temp_y/|N'|$
return (X_m, Y_m)
end for

Algorithm 3 Layer_D_Computation.

Function: Computing Delaunay Triangulation
input: Set N′
Apply *Divide and Conquer* Method
return DT

Algorithm 4 Layer_V_Computation.

Function: Voronoi Set Computation
input: Set DT
Apply *Fortune's* Method
return VP

5 Empirical evaluation

We develop the simulation scenario and implemented python script in QGIS (Quantum Geographic Information System). QGIS is an open-source desktop geographic information system (GIS) application that allows its users to visualize, modify, and analyze the geospatial information. It supports raster and vector layers both. We run it on an Intel Core 3.20 GHz machine with 8 GB of RAM running Windows 10 OS. We experimented the performance with different variations over layers and performance metrics is measured using average computation time and other statistics parameter used in our model. The map used is browsed and fetched using OpenStreetMaps.

5.1 Parameters description

Results are evaluated over different values of parameters. Table 1 highlights the brief description of the parameters used. Symbol S represents a static spatial region considered to observe the given data point distribution. Another symbol L shows the number of layers analyzed by the proposed RDV solution. P is the size of the input data point set. In our empirical evaluation it ranges from as low as 30 data points to as high as 76 data points that are scattered over different spatial region sizes namely, sizable, moderate, and the small. The review period between two consecutive run of the algorithm is taken 90 seconds.

5.2 Experimental results

Three different types of spatial region taken into consideration for the evaluation of the proposed solution. Computation time taken by the processes has no substantial impact on overall performance of RDV masking due to the fact that masked released data holding optimum privacy measures are more dominant and no real time communication is involved here. Generally, third party is outsourced to perform such computation to draw results

Table 1 Parameters used with description

Parameter	Description	Values used
S	Spatial Region Area	*Static*
L	Number of Analyzed Layers	3
P	Size of a input item	76, 49, 30
Region type	Size of Spatial Region	Sizable, Moderate, Small scale
Review period	Time interval between two consecutive run of the algorithm	90s

within a stipulated time line. Therefore, the overall execution time is measured in terms of computation complexity notation.

Different case scenarios The regions covering the given location-tagged nodes can be of different sizes. This study draws the impact of spatial region size over the pattern change behavior of the masked discrete data points. Note that for a *sizable R* the data used is of Alaska's airport sprawling over the area of 1.718 million sq.km. For a *moderate R* an area of 475 sq.km. is taken, while an area of 22 sq.km. is considered for a *small R*. Following is the description of three different regions scenarios used for analysis.

5.3 Case I: when R type is sizable

This is the case when the area covering the entire discrete data points are extremely large. We consider a country wide region and the points depicting in the results are the airport data of the region.

5.3.1 Layer R results

The Figure 16 shows the instance when the points are scattered over the plane.

In the Figure 17, the points are distributed over the region R that are fetched using OpenStreetMaps on the basis of the points given. It shows the layer R processing when the underlying map shows some characteristics related to the region's spatial degree.

5.3.2 Layer D results

Once the nodalization is achieved in layer R, layer D is computed as shown in the Figure 18. Layer D adds a delaunay triangulation layer to the region layer. Here, 76 data points are converted into series of delaunay triangles.

5.3.3 Layer V results

Layer V results are shown in the Figure 19 where every point gets separated to the closest segment along the edges of corresponding voronoi polygon.

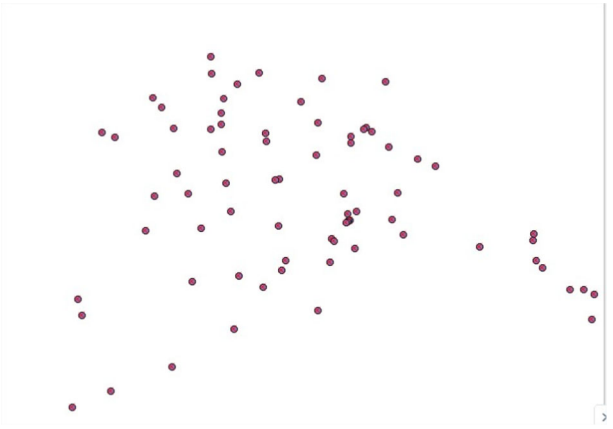

Figure 16 Scattered points

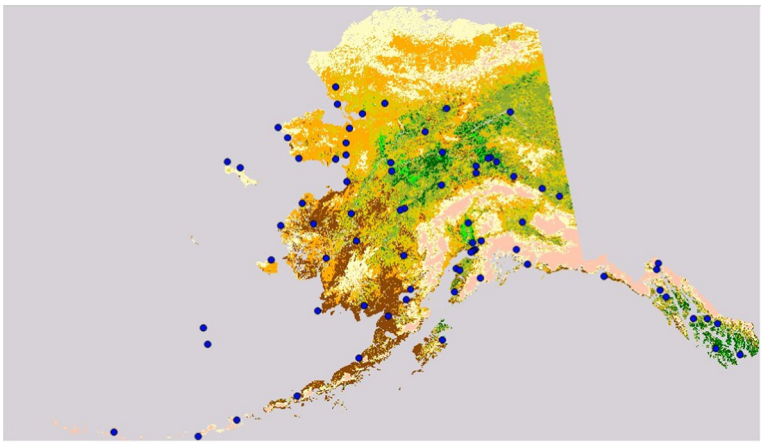

Figure 17 Point distribution over sizable region R- Part II

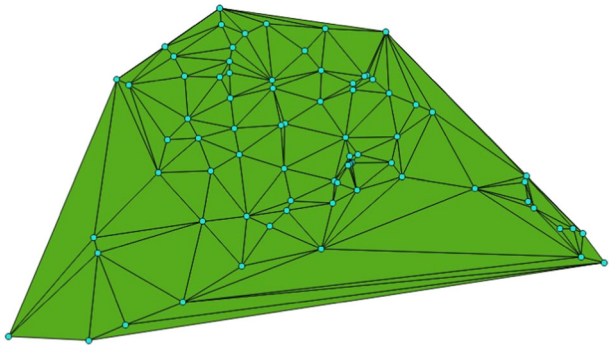

Figure 18 Computation of Delaunay triangulation under sizable R

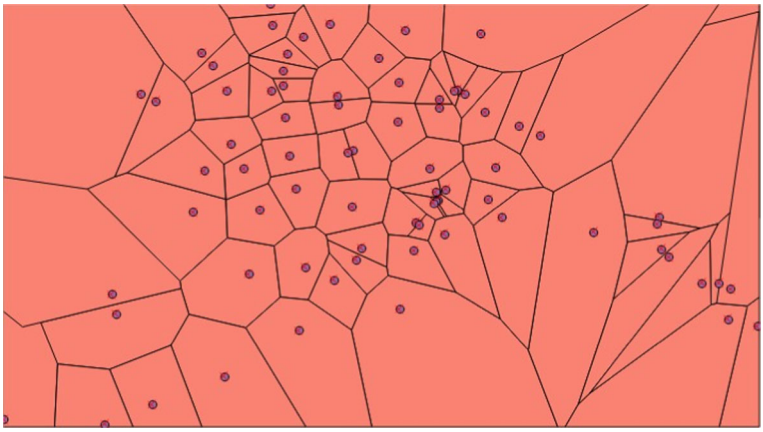

Figure 19 Voronoi polygon computation under sizable R

5.4 Case II: when R type is moderate

In this case, the points are statically chosen such that they are restricted within a moderately sized spatial region. Following are the outcomes of different computations suggested for 49 points. A city wide area containing 49 different location points is considered for the study.

5.4.1 Layer R results

The Figure 20 shows the instance when the points are scattered over the plane and the considered region is city wide where given data points represents various amenities. The size of the region is within 25 kilometers area of the city.

5.4.2 Layer D results

Layer D is computed and shown in the Figure 21. The given figure depicts the data points triangulation where every data point is at least the part of one of the produced delaunay triangles.

5.4.3 Layer V results

Layer V results are shown in the Figure 22. The voronoi polygons are then computed from the delaunay triangulation points, where the centers of the produced voronoi polygons are considered as new masked data points.

5.5 Case III: when R type is small scale

This is the case where points are statically chosen such that they are restricted within a relatively smaller sized spatial region. The following are the outcomes of different computations

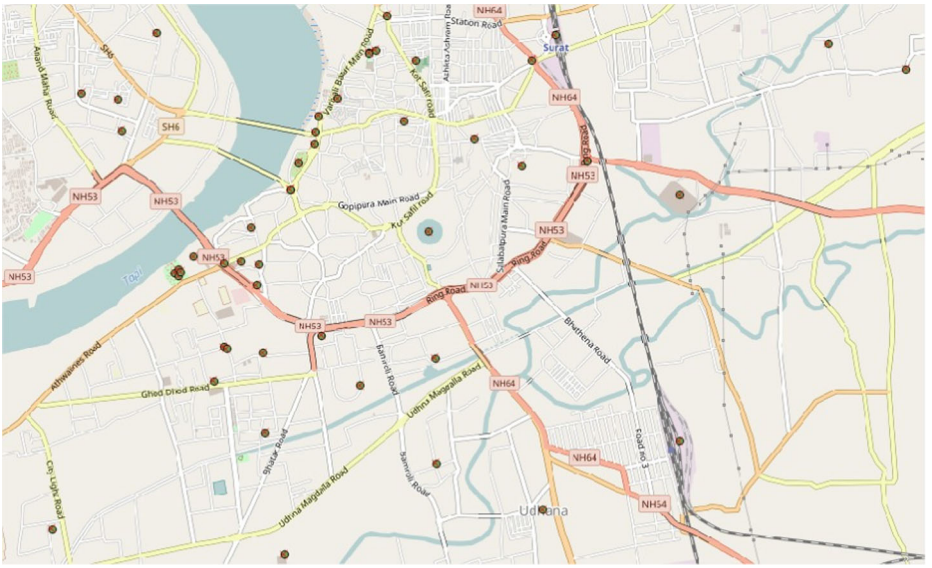

Figure 20 Scattered points with moderate R

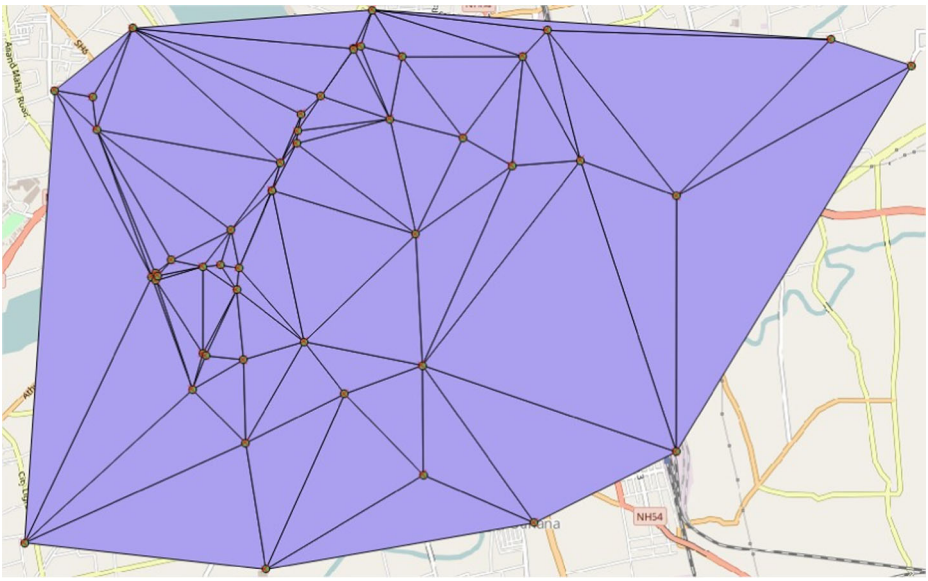

Figure 21 Computation of Delaunay triangulation under moderate R

performed for 30 points. A specific region in a city with 30 different location points are considered for the study.

5.5.1 Layer R results

The Figure 23 shows the instance when the points are scattered over the plane.

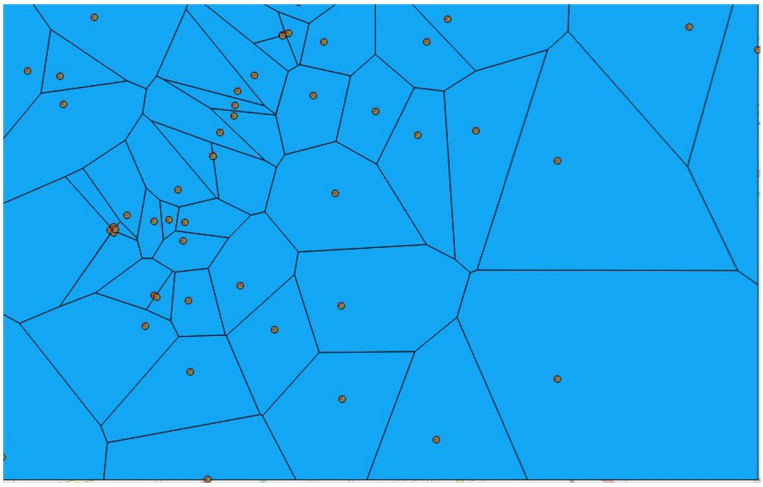

Figure 22 Voronoi polygon computation under moderate R

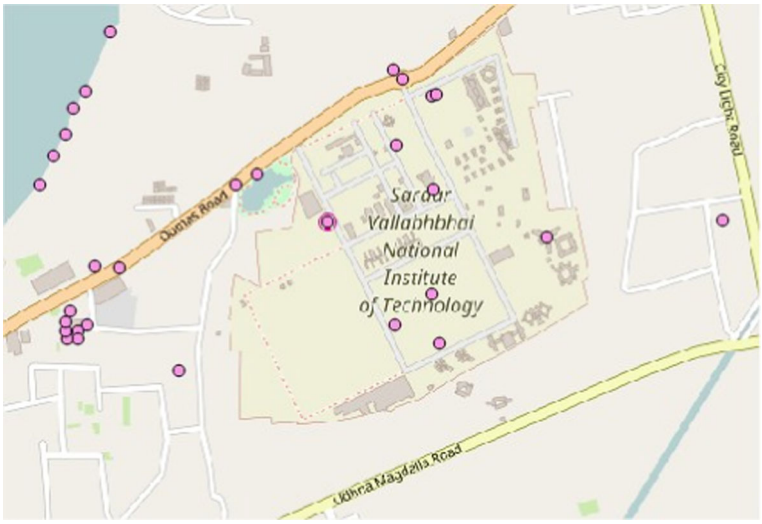

Figure 23 Scattered points with small scale R

5.5.2 Layer D results

The layer D is computed and shown in the Figure 24, that also depicts the data points triangulation where every data point is at least the part of one of the produced delaunay triangles.

5.5.3 Layer V results

The layer V results are shown in the Figure 25. Voronoi polygons are computed from the produced delaunay triangulation points, where centers of the voronoi polygons are considered as new masked data points. In small scale region the new masked data points deviates lesser distance from the original data points in comparison to the sizable and moderate R values so that the spatial data point pattern remains intact and useful for further analysis.

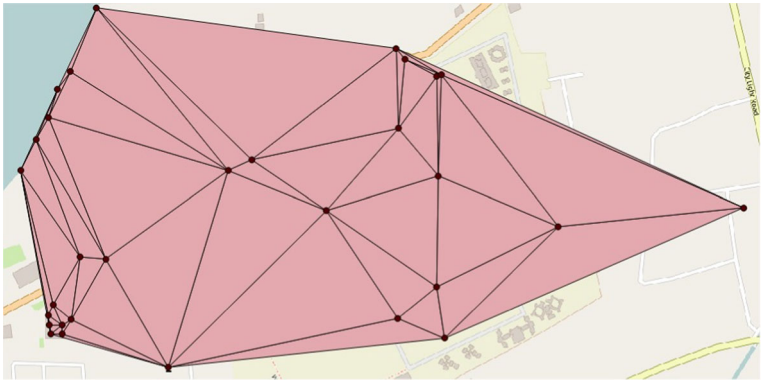

Figure 24 Computation of Delaunay triangulation under small scale R

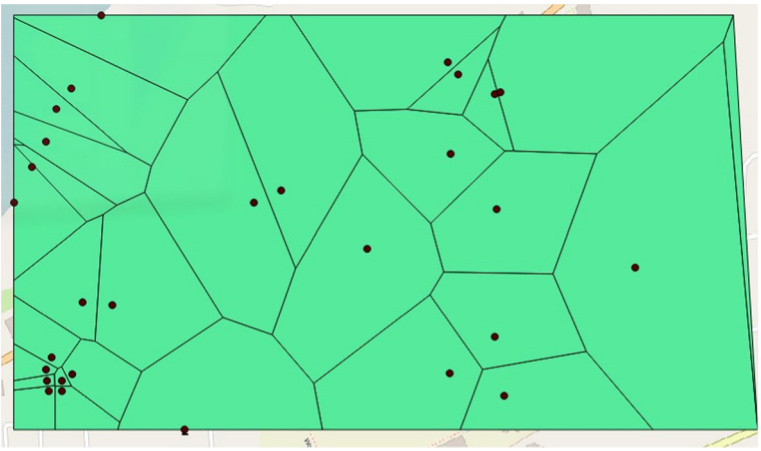

Figure 25 Voronoi polygon computation under small scale R

5.6 Nested RDV scenario

Let V be the set produced after the first round of the masking from the input point set D. Let the iterative indicator $\mathcal{I}$ ranges from $1 \ldots n$, where 1 is the lowest level (suggests only 1 iteration of RDV Masking) and n denotes the maximum level (all points are replaced by a single point as shown in the Figure 27). The Figure 26 shows an instance of iterative RDV Masking when value of $\mathcal{I} > 1$. This process is repeated iteratively for the specified value of $\mathcal{I}$.

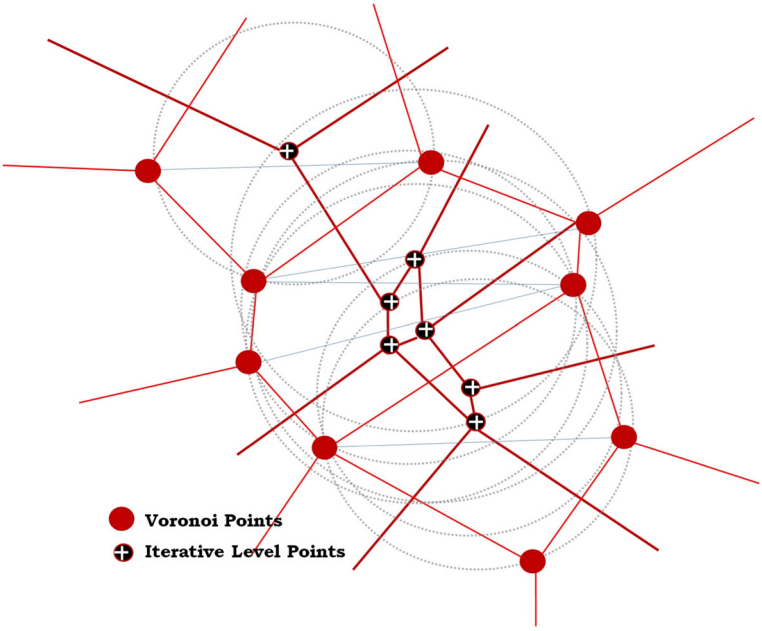

Figure 26 An example scenario

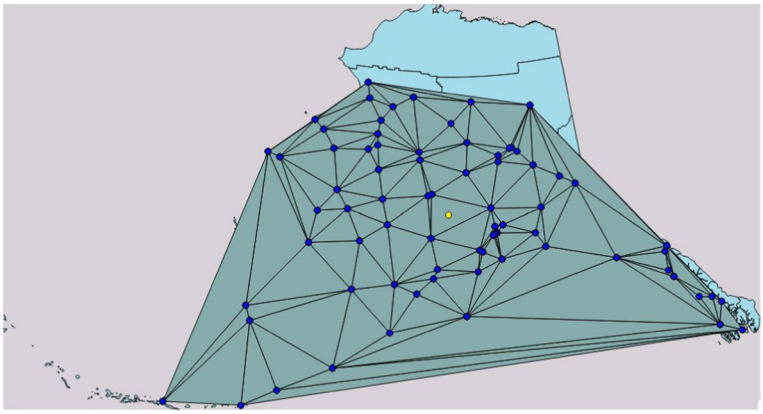

Figure 27 Maximum threshold under sizable R

5.6.1 Effect of iterative indicator

The proposed three layer RDV masking behaves differently for the different values of iterative indicator $\mathcal{I}$. We evaluate the performance of the proposed approach for the various degrees of $\mathcal{I}$ that ranges from 1 to 5, where 1 denotes the least masking degree and 5 represents the highest degree of masking. Following are the cases with various degrees of $\mathcal{I}$.

Case 1: When Maximum Threshold of $\mathcal{I}$ is selected

The Figure 27 shows the scenario when iterative indicator is chosen to its maximum value (here $\mathcal{I} = 5$) and data masking is at the most abstract level. The single yellow spot in the middle of Delaunay hill is the aggregated point that needs to be substituted against entire data points in case of the maximum threshold.

Case 2: When $\mathcal{I}$ ranges for different values

This case presents the behavior of RDV masking for different ranges of $\mathcal{I}$. The Figures 28 and 29 represents the output of layer D and layer V, respectively. This case shows the lowest

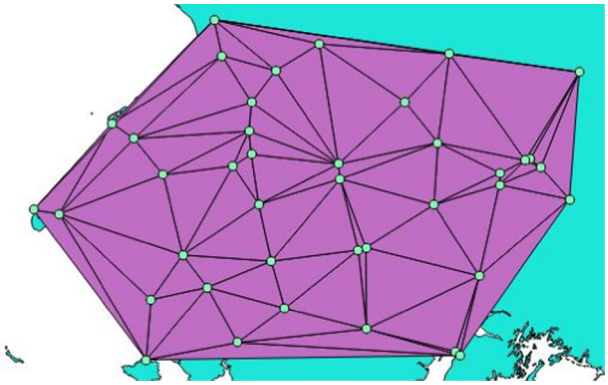

Figure 28 Impact of iterative indicator on Delaunay triangulation for sizable R when $\mathcal{I}$ =1

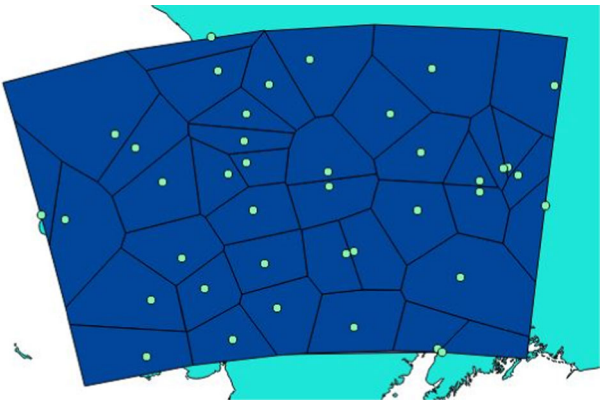

Figure 29 Impact of iterative indicator on Voronoi polygon for sizable R when $\mathcal{I}$ =1

degree of masking while the Figures 30, 31, 32, and 33 render the behavior of masking for relatively higher values of $\mathcal{I}$. The similar behavior can be generated for moderate and small sized R.

It can be observed that the spatial pattern of data points remain intact and does not deviate much even though $\mathcal{I}$ ranges within the certain specified range. Therefore, it can be noticed that the usefulness of the published data points is still preserved .

5.7 Statistical analysis

RDV masking performance is well analyzed with the attached statistical values. Table 2, 3 and 4 shows the values collected after the suggested technique.

The low standard deviation for a highly scattered given points (covering a larger spatial region) is a validating parameter to validate that the masked data points are not deviated much from the original locations, therefore, masked location data points pattern highly

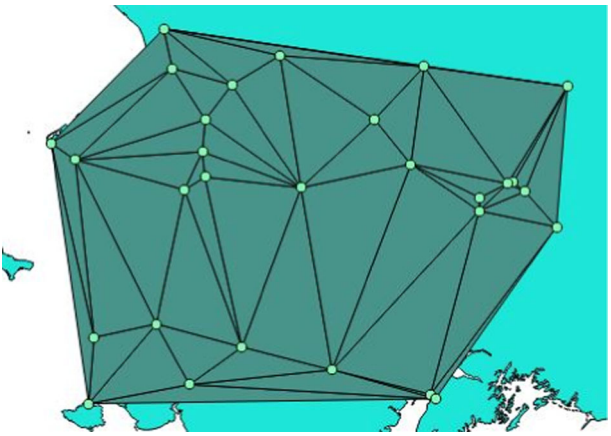

Figure 30 Impact of iterative indicator on Delaunay triangulation for sizable R when $\mathcal{I}$ =2

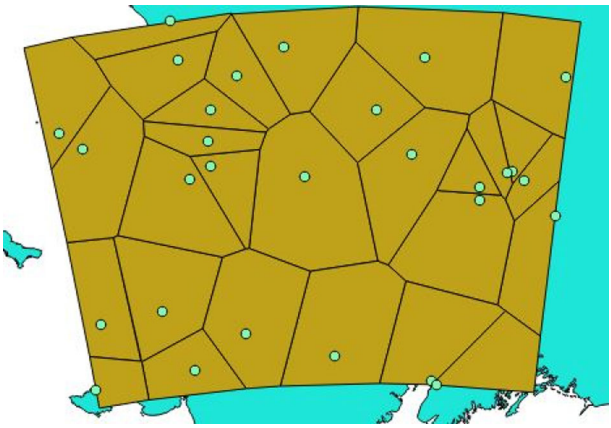

Figure 31 Impact of iterative indicator on Voronoi polygon for sizable R when $\mathcal{I}$ =2

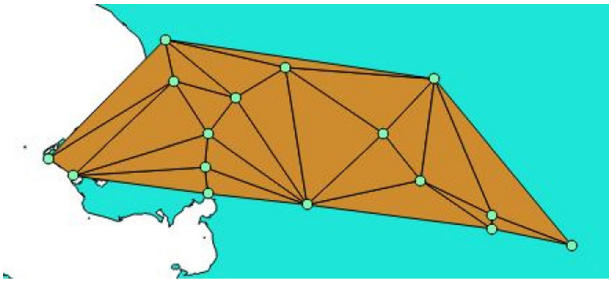

Figure 32 Impact of iterative indicator on Delaunay triangulation for sizable R when $\mathcal{I}$ =3

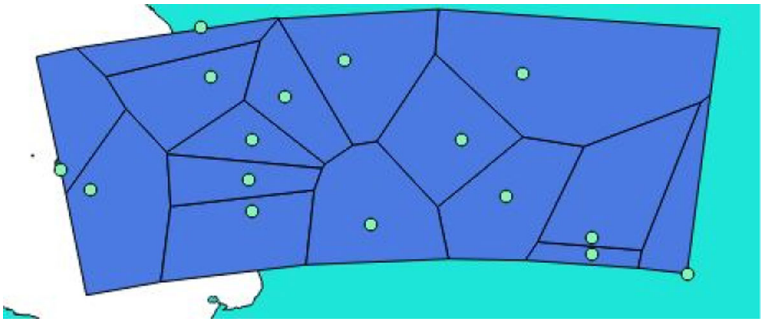

Figure 33 Impact of iterative indicator on Voronoi polygon for sizable R when $\mathcal{I}$ =3

Table 2 Statistics values table when R is sizable

Attribute	Value
Count	76
Unique values	57
Minimum value	9.0
Maximum value	1569.0
Mean value	299.447
Median value	109.5
Standard deviation	408.87
Z-Score	–3.88
Coefficient of variation	1.365

Table 3 Statistics values table when R is moderate

Attribute	Value
Count	49
Unique values	49
Minimum value	414.78
Maximum value	6730.51
Mean value	3973.19
Median value	32.08
Standard deviation	11.92
Coefficient of variation	0.336

Table 4 Statistics values table when R is small scale

Attribute	Value
Count	30
Unique values	30
Minimum value	251.50
Maximum value	763.1
Mean value	502.07
Median value	26.30
Standard deviation	2.51
Coefficient of variation	0.582

Table 5 Computational complexity

Layer R	Layer D	Layer V
O(N)	O(N log N)	O(N log N)

resembles to the original point pattern. Thus, making the released data useful for further analysis. However, this parameter is relatively high in grid masking and aggregation approaches that leads to very less or no resemblance of the masked pattern to the original data pattern which is highly unwanted for good research/ genuine analysis. Moreover, in the proposed mechanism the data points are not displaced only on the basis of density of the populace in the region but also takes other proposed iterative indicators into the account for deciding the displacement factor value that makes the restoring of the data points cumbersome for an attacker.

5.8 Computational complexity

Computational complexity of the proposed RDV masking is shown in the Table 5. It is clear that the proposed solution do not take more than O(N log N).

RDV Masking technique is able to satisfy the geomasking need of the location data and it additionally provides *iterative* masking results on the basis of other important parameters like potential user of published data, investigation type, statistics responsiveness, and degree of existed quasi identifiers. It also preserves spatial pattern of data points effectively. Other perturbation methods like random perturbation and donut masking are able to maintain the data patterns only if the perturbation factor is low such that the original point relocates over or in the close proximity of the original point. Unlike other existing mechanism our proposed three layer iterative RDV masking solution preserves privacy while also preserves spatial pattern and maintains high data utility.

6 Conclusions

The contribution of this work is two folds. First, this research study proposes a mechanism that provides an efficient way to geomasked the location attribute value before releasing the record to make it public. Second, it identifies iterative indicators (addition to the population density indicator) that attributed to the displacement computation using suggested algorithm. Our proposed algorithm; three layer iterative RDV masking exploits the basic geometry mathematics concept of Delaunay Triangulation and Voronoi Polygon formation. The major benefit of our proposed solution is that apart from providing efficacious data point masking, it also preserves the pattern of original data points. Grid masking and aggregation mechanism disrupt the original data points pattern and reduce the data utility for further research and analysis. Other perturbation methods like random perturbation and donut masking are able to maintain the data patterns only if the perturbation factor is low such that the original point relocates over or in the close proximity of the original point. Unlike other existing mechanisms our proposed three layer iterative RDV masking solution preserves privacy while preserving spatial pattern and maintains high data utility. Experimental results and privacy measurement analysis of the algorithm also support the mentioned benefits that also reduces the disclosure risk.

References

1. AbdelMalik, P., Boulos, M.N.K., Jones, R.: The perceived impact of location privacy: A Web-based survey of public health perspectives and requirements in the uk and canada. BMC Publ. Health **8**, 156 (2008)

2. Armstrong, M.P., Rushton, G., Zimmerman, D.L., et al.: Geographically masking health data to preserve confidentiality. Stati. Med. **18**, 497–525 (1999)
3. Aurenhammer, F., Klein, R.: Voronoi diagrams. Handbook Comput. Geom. **5**, 201–290 (2000)
4. Cassa, C.A., Grannis, S.J., Overhage, J.M., Mandl, K.D.: A context-sensitive approach to anonymizing spatial surveillance data. J. Am. Med. Inform. Assoc. **13**, 160–165 (2006)
5. Cignoni, P., Montani, C., Scopigno, R.: Dewall: A fast divide and conquer delaunay triangulation algorithm in ed. Comput.-Aided Des. **30**, 333–341 (1998)
6. Cox, L.: Matrix masking methods for disclosure limitation in microdata. Survey Methodol. **20**, 165–169 (1994)
7. Curtis, A., Mills, J.W., Agustin, L., Cockburn, M.: Confidentiality risks in fine scale aggregations of health data. Comput. Environ. Urban. Syst. **35**, 57–64 (2011)
8. de Berg, M., Cheong, O., van Kreveld, M., Overmars, M.: Delaunay triangulations: Height interpolation. Comput. Geom. Algor. Appl. **9**, 191–218 (2008)
9. Du, Q., Emelianenko, M., Ju, L.: Convergence of the lloyd algorithm for computing centroidal voronoi tessellations. SIAM J. Numer. Anal. **44**, 102–119 (2006)
10. Duckham, M., Kulik, L.: Location privacy and location-aware computing. Dynamic & Mobile GIS: Investigating Change in Space and Time **3**, 35–51 (2006)
11. Duncan, G.T., Pearson, R.W., et al.: Enhancing access to microdata while protecting confidentiality: Prospects for the future. Stat. Sci. **6**, 219–232 (1991)
12. Edelsbrunner, H., Shah, N.R.: Incremental topological flipping works for regular triangulations. Algorithmica **15**, 223–241 (1996)
13. Elfick, M.: Contouring by use of a triangular mesh. Cartogr. J. **16**, 24–29 (1979)
14. Elwood, S., Leszczynski, A.: Privacy, reconsidered: New representations, data practices, and the geoWeb. Geoforum **42**, 6–15 (2011)
15. Fortune, S.: A sweepline algorithm for voronoi diagrams. In: Proceedings of the Second Annual Symposium on Computational Geometry, pp. 313–322. ACM (1986)
16. Guibas, L.J., Knuth, D.E., Sharir, M.: Randomized incremental construction of delaunay and voronoi diagrams. Algorithmica **7**, 381–413 (1992)
17. Gupta, R., Rao, U.P.: Achieving location privacy through CAST in location based services. J. Commun. Netw. **19**(3), 227–238 (2017)
18. Gupta, R., Rao, U.P.: An exploration to location based service and its privacy preserving techniques: A survey. Wireles Personal Commun. **96**(2), 1973–2007 (2017)
19. Gupta, R., Rao, U.P.: A hybrid location privacy solution for mobile lbs. Mob. Inf. Syst., 2017 (2017)
20. Gupta, R., Rao, U.P.: VIC-PRO: Vicinity protection by concealing location coordinates using geometrical transformations in location based services. Wireles Personal Commun. **107**(2), 1041–1059 (2019)
21. Hampton, K.H., Fitch, M.K., Allshouse, W.B., Doherty, I.A., Gesink, D.C., Leone, P.A., Serre, M.L., Miller, W.C.: Mapping health data: Improved privacy protection with donut method geomasking. Am. J. Epidemiol. **172**(9), 1062–1069 (2010)
22. Hofmann-Wellenhof, B., Lichtenegger, H., Wasle, E.: GNSS–global navigation satellite systems: GPS, GLONASS, Galileo, and more. Springer Science & Business Media (2007)
23. Hurtado, F., Noy, M., Urrutia, J.: Flipping edges in triangulations. Discret. Comput. Geom. **22**, 333–346 (1999)
24. Kounadi, O., Leitner, M.: Spatial information divergence: Using global and local indices to compare geographical masks applied to crime data. Trans. GIS **19**, 737–757 (2015)
25. Kwan, M.-P., Casas, I., Schmitz, B.: Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? Cartographica: The International Journal for Geographic Information and Geovisualization **39**, 15–28 (2004)
26. Leach, G.: Improving worst-case optimal delaunay triangulation algorithms. In: 4th Canadian Conference on Computational Geometry, pp. 340–346. Citeseer (1992)
27. Leitner, M.: A first step towards a framework for presenting the location of confidential point data on maps results of an empirical perceptual study. Int. J. Geogr. Inf. Sci. **20**, 813–822 (2006)
28. Leitner, M., Curtis, A.: Cartographic guidelines for geographically masking the locations of confidential point data. Cartograph. Perspect. **6**, 22–39 (2004)
29. Li, M., Sun, X., Wang, H., Zhang, Y., Zhang, J.: Privacy-aware access control with trust management in Web service. World Wide Web **14**(4), 407–430 (2011)
30. Linde, Y., Buzo, A., Gray, R.: An algorithm for vector quantizer design. IEEE Trans. Commun. **28**, 84–95 (1980)
31. o'Rourke, J., Mallinckrodt, A.J., et al.: Computational geometry in c. Comput. Phys. **9**, 55–55 (1995)

32. Rebay, S.: Efficient unstructured mesh generation by means of delaunay triangulation and bowyer-watson algorithm. J. Comput. Phys. **106**, 125–138 (1993)
33. Shu, J., Jia, X., Yang, K., Wang, H.: Privacy-preserving task recommendation services for crowdsourcing. IEEE Transactions on Services Computing. https://doi.org/10.1109/TSC.2018.2791601 (2018)
34. Sinclair, D.: S-hull: A fast radial sweep-hull routine for delaunay triangulation. arXiv:1604.01428 (2016)
35. Su, P., Drysdale, R.L.S.: A comparison of sequential delaunay triangulation algorithms. Comput. Geom. **7**, 361–385 (1997)
36. Tsai, J.D.V.: Fast topological construction of Delaunay triangulations and Voronoi diagrams. J. Comput. Geosci. **19**, 1463–1474 (1993)
37. Vorono"ı, G.: Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les parallélloèdres primitifs. J. für die reine und angewandte Mathematik **134**, 198–287 (1908)
38. Wang, H., Zhang, Z., Taleb, T.: Special issue on security and privacy of IoT. World Wide Web **21**(1), 1–6 (2018)
39. Weiser, P., Scheider, S.: A civilized cyberspace for geoprivacy. In: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Privacy in Geographic Information Collection and Analysis, p. 5. ACM (2014)
40. Zandbergen, P.A.: Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. Advances in Medicine, 2014 (2014)
41. Zhang, J., Tao, X., Wang, H.: Outlier detection from large distributed databases. World Wide Web **17**(4), 539–568 (2014)
42. Zhang, J., Li, H., Liu, X., Luo, Y., Chen, F., Wang, H., Chang, L.: On efficient and robust anonymization for privacy protection on massive streaming categorical information. IEEE Trans. Depend. Sec. Comput. **14**(5), 507–520 (2015)