Do You Know Where Your Cloud Files Are? Improving Accuracy of Cloud Location Detection using Modified K-Medoid Algorithm

Kalpit G. Soni* and Atul Patel

Charotar University of Science and Technology, Changa - 388421, Gujarat, India; kalpitsoni1@yahoo.com, atulpatel.mca@charusat.ac.in

Abstract

Background/Objectives: The emergence of Cloud computing in recent times as a globally accepted data and activity management platform is phenomenal, albeit with few clutches not only related with the data integrity and their privacy but also of their geographical location. The gravity of the issue arises with the plethora of guidelines and variations in legal procedures in storing the data and getting access to them in diverse geographical areas. The aim of present investigation is to figure-out a potentially effective method that can help detect a forged server with better location coordinates than the currently employed methods. Methods/Statistical Analysis: In this study we developed a made use of an established method to check the position of the files at a specified place as claimed by service provider. The main means employed during the course of this investigation is K-Medoid Algorithm for finding the server’s geographical location putting emphasis on putative adverse outcome. Findings: The greatness of K-Medoid Algorithm is to detect a potentially forged location and also to identify the possible elements involved in forgery. This work involved the use of Tulip Dataset and Implementation in NS3 for validating algorithm modified by us to detect the forged location and possible location of involved adversaries. The investigation during this work culminated in proposing a modified K-Medoid algorithm for helping a user to find the Cloud file system which verifies the storage location with very high efficiency. Improvements/Applications: Through our research technique, user able to detect faked location and identify the dishonest service provider. Our work culminated in the best estimation of the location of server for a cloud service provider within a less distance than the previous work.

Keywords: Cloud Computing, Geolocation, K-Medoid, Trilateration

1. Introduction

There are many methods for managing and storing data and activities on various platforms. Among these, the cloud computing is gaining favor towards a universally accepted data and activity management platform. It is increasingly gaining favor towards a universally accepted data and activity management platform. The global impetus towards cloud platform ushered when the US federal government announced its “Cloud First” policy to plan migrate about 75% of their data management task on Cloud in order to comply to their policy. Moreover, small and big, corporate houses have started renting or developing their own cloud platforms for transferring their data processing activities. The Users in this mode give up their control on data in a bid to get faster-on-demand resources and to share the administrative costs once they transfer off their data from traditional servers to Cloud. The trust and dependency on the Cloud service provider is an inherent demand of this transition. The sanctity of one's data is the most important criteria for this revolution which is marred by recent controversies when Hotmail user's data were wiped out and also the permanent loss of data suffered by Amazon’s EC2 Cloud service. In Cloud Service Platform are an exclusively cheap and efficient way to manage data but with an element of security and breach of conduct.
2. Geolocation Background

Numerous services, restriction, fraud detection systems contribute to identifying the geographical position of Internet host. In the current era, identifying the position of Internet host has been a significant problem to deliberate.

The importance of geographically locating the hosts on the World Wide Web (WWW) is enormous for many applications which use IP Geolocation. For example online entertainment providers like⁴,⁵ have resorted to keeping their content tailored according to the location of their clients. IP geolocation aims to solve the hassle of figuring out the geographic location of a given IP address to pin-point on the exact city or country location of the IP address which can be scaled efficiently in the longitude and latitude coordinates of the target.

2.1 Scope

Broadly, the problem is to actively monitor if a storage service provider is meeting its geographic obligations. Our scope is verifying the position of known data copies. Consequently, it is important that a client is able to verify that the service paid for are the services being provided. One approach to such verification is to rely on contracts or Service Level Agreement (SLA). However, a contractual approach may not detect misbehavior – whether malicious or accidental –It is better to augment such an approach with a mechanism by which clients can themselves verify that the properties they depend on the certain preserve. Many aspects of cloud computing can be verified by the end user. The advantage with cloud computing is freedom to users and service providers to place their storage at any specified location. The users have the freedom to store the files at a region of their choice and later on they can verify whether the files are stored properly at their chosen location since the importance is to be sure of the location of the file. This becomes more important in the scenario where users may take legal actions citing the violation of SLA if they found that files are not stored properly in their requested region. The modus-operandi of adversaries lies in perturbations in delay time; whether shortening or lengthening. To detect the location of forgery and zeroing in on specific servers we employed Trilateration technique⁶. Trilateration involves the use of an ample number of distances to few fixed points for the purpose of estimation. The Landmarks acting as fixed or reference points were used to send challenges to the Target servers, which was our point of interest.

2.2 Goal

The aim of this study culminating in verification of a storage location and detection of the forgery points in a cloud-based platform. If there is a violation of SLA guidelines by service provider as observed by the user in the form of invalid location of storage then they are able to prove it and take legal actions against them.

Figure 1 describes a typical case of ideal file storage Cloud platform with honest service provider. Here the user has the ability to ascertain the geographical location of their files (f) stored by the service provider through standard Auction process⁷ i.e., selection of a group as per Cloud service provider's criteria, in the servers S1 and S2 in the R₁ Region. The user can find out the approximate geographical location by calculating Round Trip Time (RTT) sending challenges through some Landmarks of trust. Hence, the user can stay sure about the sanctity and position of their stored data along with a safeguard against potential hack in the form of position forgery.

2.2.1 Model-1, Honest Service Provider with Unintentional Error

Figure 2 describe a situation where service provider is honest and had no vested interest of fiddling with User’s data but an inadvertent error in the protocol or a tangible infrastructure led to the storage of file (f) in the Region R₁ as requested by User but in a server other than requested. The queries are sent by user to service provider for identification of servers storing the files. Suppose if s₁ and s₂ were to store the file but by mistake, the file was stored...
in the server s6 rather than in s2, then user will get an improper response from s2 on sending a query to s1 & s2 using a Landmark that sends a legal challenge.

**Figure 2.** Honest service provider with unintentional error and intentional modification.

### 2.2.2 Model-2, Honest Service Provider with Intentional Modification

This model as explained by Figure 2 poses a situation where the service provider keeps the files in the location; s6 instead of s2, which is either cheap or convenient for them but no intentions of a data breach. When a user sends the challenge for the location of the file to the server, he comes to know that the reliability test of the server is passed but since RTT is altered hence the geographical location of the server is different than what is expected.

### 2.2.3 Model-3, Dishonest Service Provider Limited by “Laws of Physics”

This model explains a condition where the service provider is not honest in disclosing the position of its server. They store files at other geographical location than promised. T\textsubscript{\text{Real}} is the place of storage of file as per the agreement between Cloud service provider and the User, but the service provider deceives users to place files at T\textsubscript{\text{Fake}}, due to pricing or legal factors. Deliberate delay in RTT introduced by service provider due to geographical vicinity can’t be ascertained when the challenge was fired by the user from LM\textsubscript{1}. But since shortening of RTT in case challenge was sent from LM\textsubscript{2} & LM\textsubscript{3} is not possible according to limit of the speed of light (C= 3 X 10\textsuperscript{8} m/s) hence service provider is obliged to locate the server at a preplanned position of T\textsubscript{\text{Plan}}, which is near T\textsubscript{\text{Real}}. However, T\textsubscript{\text{Real}} does not suffice with the cost or geographical requirements of the service provider hence it’s very improbable that service provider will place the servers at T\textsubscript{\text{Plan}} which are very near to T\textsubscript{\text{Real}} as shown in Figure 3.

**Figure 3.** Dishonest service provider limited by “Laws of Physics”.

### 2.2.4 Model-4, Dishonest Service Provider with no “know-how”

In this mode of providing service, the Cloud service provider would need to add delay to RTT due to placing the server in the shorter range from the user as promised. For that the service provider places an operating machine near T\textsubscript{\text{Real}} to send a challenge from this position to Landmark in a bid to ascertain the amount of delay time to be added, but in such a case there is always an element of doubt that can be traced back to the faked position of server because this machine is not in the list of trusted machines of the user. Hence, not knowing the exact value to delay time to add, the service provider uses the limitation of network traffic speed corresponding to 66% of the velocity of light. In the essence of present work as explained in this paper strives to explain the Model-4, using K-Medoid algorithm to figure-out the potential discrepancies in the location of the server as claimed by the service provider and to check the location of the files. K-Medoid algorithm is a very effective method to find out the authenticity of service provider’s claim about the geographical location of its server.
3. Location Identification Method using Landmarks

The desired IP’s are located by calculating the delay in RTT by using algorithms meant for delay-based geolocations. The present study used Trilateration method to transform the measured delay in time into the physical identification of servers putting to use the methods as explained by Gueye, B., et al., which involve Constraint-Based Geolocation of hosts. In this method, each Landmark is required to send ping other Landmarks placed at appropriate distances and the delay time generated is identified. Then, a graph of delay in RTT is plotted versus distance between Landmarks as explained in Figure 4; the graph hence generated is a straight line curve that includes all the data points which are close but above the line as explained by straight line equation.

\[ y = mx + c \]

Figure 4. RTT (ms) Vs distance (KM) calibration.

In this equation, \( y \) represents a delay in RTT in milliseconds and \( x \) represents the distance between Landmarks in kilometers. \( M \) represents the slope of the graph. Where \( M \) is calculated using slope \( M = \tan \Theta = \text{RTT}/\text{Distance} \). It shows longer in delay time; higher will be the value of \( \tan \Theta \). \( c \) represents intercept, i.e., that value of RTT when the distance between Landmarks is supposed to be zero. It represents the innate time delay in transmission of the message.

The most appropriate line lying below actual data points for individual Landmarks (\( L_i \)) involving its intercept is represented by

\[ y_i = mx_i + c_i \]

Although this most appropriate line is a different entity then the line of best fit curve, as obtained by doing single variant linear regression analysis and it lies below best fit curve. The most appropriate line obtained for each Landmark can be used transform the delay time into the geographical distance by projecting a perpendicular from most appropriate line on to the distance axis of the graph albeit with a fact that it is going to give longer than actual distance since it is not a line involving real data points.

The Landmark acts as the Centroid of the circle that is being constructed using its distance obtained from most appropriate line as the radius. This makes sure for each Landmark to have the sought-after server near its periphery. Since numbers of Landmarks employed in this study are three hence the total number of curves generated is also three. The point of crossing of each curve encloses a region that gives a set of information for the exact location of the sought-after server. The individual points are measured by drawing a Centroid of the area enclosed by points of crossing of each curves having vertices as explained in Figure 5 by using the formula.

\[
C_x = \sum_{i=0}^{n-1} (x_i + x_{i+1}) (x_i/x_{i+1} \quad y_i/y_{i+1})
\]

\[
C_y = \frac{1}{6A} \sum_{i=0}^{n-1} (y_i + y_{i+1}) (x_i/x_{i+1} \quad y_i/y_{i+1})
\]

Figure 5. Trilateration for geolocating.

4. Detecting Location Forgery

Figure 6 explains a situation where the service provider is dishonest in revealing the exact position of its servers and strives to make users feel the server is a \( T_{\text{Real}} \) but in actuality, it is at \( T_{\text{Fake}} \). The curves formed by three Landmarks
selected are r1, r2 and r3. But since the dishonest service provider wishes to hide the location of server, it tries to shift the Centroid of the region formed at the point of crossing of each curve; r1, r2 and r3 towards T Fake considering the limitation posed by speed of light and hence estimating the delay required to be added to each of the Landmarks LM1. But that leads to increasing the radius of the curve formed by LM1 while those of Landmarks LM2 and LM3 will remain same. It will lead to the difference in distance between T Real and T Plan and the difference of angle between them is Θ. However, the region with increased are as enclosed by new Centroid covers radiuses r_1, r2 and r3 may warn the user about faked location of server hence the dishonest service provider would try best to reduce the angle Θ in order to not get caught.

5. System Evaluation

Firstly, the delay-to-distance was calculated as the most appropriate line. Then from three out of total 11 nodes chosen, challenges were sent to anyone of them that acted as sought-after server which resulted in the time delay for each Landmark and the concerned sought-after server. These time delays were then converted into geographical distances to make circles about them. The area enclosed by all the involved servers gave an area with a strong probability of sought-after server with Centroid representing its probable location. Once the area calculated crosses the threshold value it signifies the forgery in the location of the placement of server.

6. Details of the Experiment

Out of total 11 nodes selected from Tulip dataset, two were taken each as fake and real servers. Each of the nine Landmarks was employed to ping the rest of eight Landmarks, to generate RTT. Haversine Formula was used to calculate the geographical distances. This formula uses longitude and latitude values between two points placed on a sphere to draw a scattered plot of Distance with RTT. The server that was considered real was not used to calculate this graph.

\[
\begin{align*}
A &= \sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) \\
B &= \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right) \\
d &= 2r \arcsin\left(\sqrt{A + \cos(\varphi_1)\cos(\varphi_2)B}\right)
\end{align*}
\]

Figure 6. Location forgery.

Figure 7. Target server (LM_10) using tulip dataset.

Figure 8. Centroid using trilateration.
Figure 7 shows a sought-after server; Certennial_United State (LM\textsubscript{10}) with Landmarks surrounding it while the trilateration server used for sending challenges included; Saskatoon_Canada (LM\textsubscript{1}), Santa Cruz_United State (LM\textsubscript{2}) and Clemson_United State (LM\textsubscript{3}). The required circles were drawn using MATLAB employing the latitudes and longitudes for forming the axis of the graph. The overlap of the region of various circles formed gave a rough approximation of the location of Certennial_United State as shown in Figure 8.

Out of C\textsubscript{3} \( \binom{9}{3} \) i.e. total 84 possible combinations of nine nodes, three nodes that were chosen for sending challenges. In the set D, each data points were used to calculate the distance between them and all other data points. The pair closest to D was generated and stored in set Aj (1 \( \leq i \leq k \)). The set Aj containing these two data points were removed from Set D. The average distance between these two data points were used to measure the shortest distance from every other data points set D, the data points in set D which were nearest to average of these two points were again added to Aj and then removed from set D. This process was repeated until the number of data points in Aj reached a limit as set K-Medoid algorithm equal to \( \infty \times \frac{n}{k} \), where \( \infty = 0.7 \).

Then the Centroid was calculated using average of each.

After completing this process, initial Centroid using arithmetic mean \( C_i \) of the vectors of data points in Aj was calculated. After generating initial Centroid (\( C_1 \), \( C_2 \) and \( C_3 \)), the distance between data point to all these Centroid and associate each data point the nearest medoid was measured. Then was calculated the entire cost of the configuration by swapping data point m and o for each medoid m and for non-medoid data point o as shown in Figure 9. This process was repeated until there was observed no variation in medoid.

7. Result

We have implemented previous work with our dataset; the estimated distance is around 65 KM. The output of this research work led to an estimated distance of 54.0503 Km circle under which the location of server placed by the dishonest service provider is certain as shown in Figure 10. This implies using our method a dishonest service provider breaching the legal contract with user can be located with better precision than the latest work.

8. Conclusion

The investigation during this work culminated in proposing a modified K-Medoid algorithm for helping a user of a Cloud file system in verifying the storage location with very high efficiency. During this work, both honest and dishonest Cloud service providers were analyzed using modified K-Medoid algorithm and came up with the conclusion that this algorithm is extremely successful in zeroing in on dishonest service provider. Our scheme provides technique through which user will be able to detect faked location and identify the dishonest service provider. Our work culminated in the best estimation of the location of server for a cloud service provider within a distance which is almost 11 Km lesser than the previous work. Hence K-Medoid algorithm can safely be used to take prerogative action in case of breach of contract by service provider.
9. References

2. Murphy D. Hotmail users report blank inboxes.
3. Blodget H. Amazon’s cloud crash disaster permanently destroyed many customers’ data.
5. Sling TV. Available from: https://www.sling.com/