Placing Landmarks to Locate Internet Hosts

Artur Ziviani[1,2], Serge Fdida[1], José F. de Rezende[2], and Otto Carlos M. B. Duarte[2]

[1] Laboratoire d'Informatique de Paris 6 (LIP6)
Université Pierre et Marie Curie
Paris, France

[2] Grupo de Teleinformática e Automação (GTA)
COPPE/EE – Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil

Abstract

The deployment of a geographic location service for Internet hosts enables a whole new class of location-aware applications. We focus on a technique that provides host locations using network delay measurements from geographically distributed landmarks (hosts with a known geographic location). The problem we deal with is where to place such landmarks. We propose a demographic landmark placement approach to improve representativeness of each landmark with respect of the hosts to be located. The results show that a relatively small number of landmarks are sufficient to cover, within a limited range, the most part of hosts to be located. For a fixed number of landmarks to be placed, the demographic landmark placement provides locations to host landmarks reducing the distances from the most part of hosts to the nearest landmark.

1 Introduction

Knowing the geographic location of an Internet host from an identification of that host (name or IP address) enables a whole new class of location-aware applications. The concept of location-awareness is fundamental to pervasive computing [1] and allows the development of location-based services that take into account where the users are accessing from. Examples of such new Internet services are: localized advertising allowing publicity with regional interests on web pages, automatic selection of a language to first display the content, location-based accounting of users to classify the incoming users based on where they are, restricted content delivery following regional restrictions, or location-based security checks to authorize transactions only when performed from pre-established locations.

This work was sponsored by CAPES/COFECUB, FUJB, and CNPq.
Artur Ziviani has a scholarship from CAPES/Brazil.

Novel location-aware services take advantage of knowing the geographic location of hosts instead of only relying on the identifications of hosts. In the Internet context, however, there is no direct relation between identifications of hosts and their physical locations. A DNS-based approach was proposed in RFC 1876 [2]. Nevertheless, the DNS-based approach is not widely adopted since it requires changes in the structure of DNS records and administrators have no motivation to register location records. Alternative techniques [3] have been proposed to infer the location of an Internet host from the DNS names, from clustering the IP address space, or from measurements.

The first technique tries to infer the location of a host based on the DNS name of the host or other nearby network node. Such a technique is the base of NetGeo [4] and GeoTrack [3]. Quite often network operators assign names to routers that have some geographic meaning, presumably for administrative convenience. For example, the name `bcr1-so-2-0-0.Paris.cw.net` indicates a router located in Paris. However, not all names contain an indication of location. Moreover, as there is no standard, commonly operators develop their own rules for naming their routers even if the names are geographically meaningful. Therefore, the parsing rules to recognize a location from a node name must be commonly ISP-specific. The creation and management of such rules is a challenging task as there is no standard to follow. As the position of the host to be located is estimated as the position of the last recognizable router in the path towards the target host, a lack of accuracy is also expected.

The second technique operates splitting the IP address space into clusters such that all hosts with an IP address within a cluster are likely to be co-located. Knowing the location of some hosts in the cluster and assuming they are in agreement, the technique infers the location of the entire cluster. An example of such a technique is GeoCluster [3]. However, the technique relies on information that tends to be partial and possibly inaccurate. The information is partial as it comprises location information for a relatively small subset of the IP address space. Moreover, such an information may be inaccurate as the databases rely on data provided by users which may not necessarily provide accurate information.

The third technique relies on delay measurements and the exploitation of a possible correlation between geographic distance and network delay. GeoPing [3] is based on such a technique. The correlation between geographic distance and network delay if considered worldwide is poor, although it tends to be stronger in specific regions having richer connectivity (see Section 2 for experiments). The location estimation of a host is based on the assumption that hosts with similar network delays to some other fixed hosts tend to be located near each other. Therefore, given a set of landmarks (hosts with a known geographic location), the location estimation for a target host is the location of the landmark presenting the most similar delay pattern to the target host. The accuracy of the final location estimation and the network load due to measurements depend basically on the number of landmarks and on where they are located.

Given a finite number of landmarks, we are interested in where to place them to maximize the overall performance and the host representativeness of each placed landmark. Landmarks are expected to reflect where the most part of users (and hosts) are, thus providing more accurate location estimations for the majority of users. We propose and evaluate the demographic landmark placement that takes into account the geographic distribution of users (and consequently of hosts to be located) to place landmarks. We investigate two approaches to handle the demographic landmark placement issue. The first approach considers maximizing covered hosts for a fixed coverage distance and limited number of landmarks. The second approach considers fixing the number of landmarks while minimizing the maximum coverage distance from users and the nearest landmark.

Results show that the demographic landmark placement provides a relatively small number of landmarks able to represent a large portion of users within a limited coverage distance. Adopting fewer landmarks to locate a host implies in a lower amount of measurement traffic. We also verify that for a fixed number of landmarks, the demographic landmark placement improves the representativeness of each landmark. Such an improvement in representativeness results in closer landmarks and more accurate location estimations for the most part of hosts to be located.

The remainder of the paper is organized as follows. In Section 2, we study the correlation between geographic distance and network delay. A formalization of the host location based on inference from delay measurements is presented in Section 3. Section 4 contains the proposition of demographic landmark placement and its evaluation. Related work and further applicability of the proposed methodology are discussed in Section 5. Finally, in Section 6 we present our concluding remarks and perspectives for future work.

2 Correlation between Distance and Delay

In order to study the correlation between geographic distance and network delay, we perform measurements from the LIP6 laboratory located in Paris, France, to 135 target hosts with well-known locations all over the world. The set of target hosts are mainly composed of university sites extracted from library servers around the world [5]. The distribution of target hosts is as follows: 55 in North America, 44 in Western Europe, 7 in Eastern Europe, 11 in Asia, 7 in Latin America, 3 in Oceania, 3 in Africa, and 4 in the Middle East. From the 135 original target hosts, 109 hosts (approximately 80%) have answered the `ping` requests. The considered delay is the minimum of several measurements to avoid considering delays due to congestion on intermediate routers.

Figure 1 presents the scatter plot of the geographic distance and the considered minimum delay between Paris and each target host. We observe a weak correlation ($R^2 = 0.2971$) between geographic distance and network delay worldwide. Moreover, some points are significantly away from the remaining observations. For about the same geographic distance, the observed delay may be greater by one order of magnitude. For example, points 1, 2, and 3 in Figure 1 correspond to hosts located in Algeria, Turkey, and Iran, respectively. The `ping` packets from Paris to Algeria in the case of point 1, for instance, actually make their way through routers located in New York. The route taken towards some locations may take directions far from the straightforward physical direction (geographic properties of Internet routing are studied in further detail in [6]). Such inconvenient is strongly reinforced due to the poor connectivity in certain regions.

Poor connectivity weakens the correlation between geographic distance and network delay. From the set of 109 answering hosts, the 80 hosts located in North America and Western Europe have been identified. Such regions are likely to have the richest connectivity linking their hosts. Figure 2 shows the correlation between geographic distance and network delay considering only the hosts located in North America and Western Europe.

The regions likely to have a rich connectivity have a much stronger correlation between geographic distance and network delay. Even if there are a few isolated points away from the remaining ones, we clearly identify two main clusters in Figure 2. Such clusters correspond respectively to hosts located in Western Europe and North America. Excluding

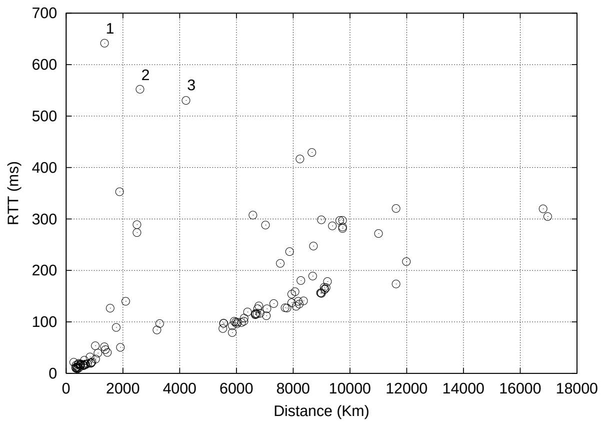


Figure 1: Scatter plot of geographic distance and network delay.

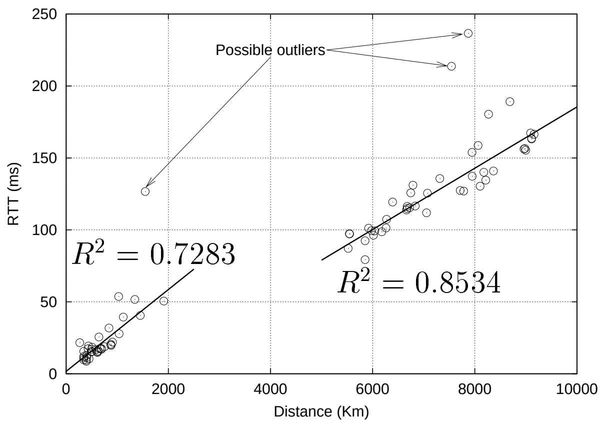


Figure 2: Correlation between distance and delay (richer connectivity).

the points that clearly remain out of the pack for both clusters (one outlier for the Western Europe cluster and two outliers for the North America cluster), we get the following degrees of correlation: $R^2 = 0.7283$ for the Western Europe cluster and $R^2 = 0.8534$ for the North America cluster.

From this simple study we observe a weak correlation between geographic distance and network delay worldwide. Nevertheless, a much stronger correlation is observed on regions presenting richer connectivity, indicating that the correlation gets stronger as connectivity gets richer. Similarly, within regions having richer connectivity, hosts located in nearby areas have similar delays to the fixed point of measure.

3 Inferring Host Location from Measurements

In the general case, there is no strong enough correlation between distance and delay to capture such a relationship under a mathematical model. Padmanabhan *et al.* [3] propose GeoPing, which adopts an empirical approach based on the observation that hosts sharing similar delays to other fixed hosts tend to be near each other geographically.

Here we formalize the problem of inferring a host location from measurements. Consider a set $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ of K landmarks. Such set constitutes the reference landmarks, which are hosts having known geographic location. Consider a set $\mathcal{P} =$

$\{P_1, P_2, \ldots, P_N\}$ of N probe machines ($N < K$). The probe machines determine periodically the network delay (actually the minimum delay value from several measurements) to each landmark. Therefore, each probe machine P_x ($x = 1, \ldots, N$) keeps a delay vector $D_x = (d_1, d_2, \ldots, d_K)$.

Suppose we desire to determine the location of a given target host T. We contact a location server that knows the set of landmarks $\mathcal{L}$ and of probe machines $\mathcal{P}$. The location server asks the N probe machines to measure the delay to the host T. Each probe machine P_x ($x = 1, \ldots, N$) returns a delay vector $D'_x = (d_1, d_2, \ldots, d_K, d_T)$, that is, the delay vector D_x constructed offline plus the just measured delay to the host T. After receiving the delay vectors from the N probe machines, we are able to construct the delay matrix $\mathcal{D}$ with dimensions $(K+1) \times N$:

$$\mathcal{D} = \begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1N} \\ d_{21} & d_{22} & \ldots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{K1} & d_{K2} & \ldots & d_{KN} \\ d_{T1} & d_{T2} & \ldots & d_{TN} \end{bmatrix} \tag{1}$$

It should be noted that the delay vectors gathered by the demanding location server from the probe machines correspond to the columns of delay matrix $\mathcal{D}$. The location server then compares the lines of the delay matrix $\mathcal{D}$ to estimate the location of host T. To infer the location of the host T, the nearest landmark with respect to T in the N-dimensional delay map is determined. Euclidean distance is used to measure the distance of the delay vectors of the K landmarks to the delay vector of host T. The landmark L which provides the smallest distance $d'_y = \sqrt{(d_{y1} - d_{T1})^2 + (d_{y2} - d_{T2})^2 + \ldots + (d_{yN} - d_{TN})^2}$, where $y = 1, \ldots, K$, is the nearest landmark with respect to host T. The corresponding location of the nearest landmark is the location estimation of the host T. The delay matrix $\mathcal{D}$ combined with the knowledge of the location of the landmarks of the set $\mathcal{L}$ compose a delay map recording the relationship between delay and location.

Let Δ denote the period of time adopted by the probe machines to periodically measure the delay to the landmark set $\mathcal{L}$. The total number of messages N_{msg} to estimate the location of h different hosts in a time interval τ is

$$N_{\text{msg}}(h, \tau) = 2N\left(\left\lceil \frac{\tau}{\Delta} \right\rceil K + h\right). \tag{2}$$

Knowing the set of adopted landmarks allows to add to the location answer indications on the possible accuracy associated with the answer based on the knowledge of the physical distribution of landmarks. For example, an estimation indicating that the nearest landmark with respect to a given host T is a landmark located in Rio de Janeiro, Brazil, should provide as estimation South America if it is the only landmark in the continent, and not Rio de Janeiro. The accuracy depends basically on the number of landmarks and on where they are located. Therefore, we are interested in where to place a finite number of landmarks to maximize performance and representativeness of each placed landmark.

4 Demographic Landmark Placement

A landmark is a reference to be used as a location estimation to a certain quantity of hosts supposed to be nearly located. As a landmark is a reference for the geographic position of a set of nearby hosts, ideally landmarks should represent the position of as many hosts as possible. High density areas in terms of hosts having a co-located landmark provide location estimations that reflect more accurately the positions of a larger number of hosts. Therefore, landmarks are expected to reflect where the most part of users (and hosts) are. We define the placement of landmarks while taking into account the user population (and consequently of hosts) distribution as the demographic landmark placement. Recent findings [7, 8] indicate a strong correlation between population and router density in economically developed nations and in urban areas characterized by population density peaks.

We first consider the main urban agglomerations spread worldwide as they are likely to offer the highest concentration of users (and as a consequence hosts to be located). We consider all urban agglomerations with more than one million inhabitants [9], totalizing 407 locations. It is known that Internet infrastructure varies dramatically across different regions throughout the world. Therefore, we weight the populations of the different agglomerations with the number of on-line users in the country the agglomeration belongs to over the total population of the country. In applying such a weight, we estimate the main user agglomerations worldwide and their demands to be covered by the demographic landmark placement. Data on estimations of the on-line and total population of each country are available at [10].

The possible locations of the landmarks are limited to the locations of the agglomerations. The solution to the placement problem under such a restriction may be inferior to the solution to the problem if the landmarks are allowed to be located anywhere. Nevertheless, landmarks are hosts likely to be unaware of being used as landmarks. Therefore, it is more suitable to place landmarks where is likely to exist previous network infrastructure, as in the case of the user agglomerations.

Eligible landmarks are any host able to echo `ping` messages and known to be located within the indicated urban agglomerations, thus resulting in a potential large number of eligible landmarks. Thus, for each considered urban agglomeration a set of hosts may be selected to constitute a pool of eligible landmarks, which could even be unsuspecting participants in the procedure. As probe machines need to use one landmark from a certain urban agglomeration, such probe machines pick one of the eligible landmarks located in the urban agglomeration. The association between a geographic location and the pool of eligible landmarks located therein may be performed using DNS, performing load balancing among co-located landmarks. Each pool of landmark machines could be named `new-york.netlandmark.info`, `london.netlandmark.info`, and so forth rather than adding IP addresses into the code of probe machines, thus enhancing robustness against individual landmark failures.

Denoting the set of user agglomerations (the set of candidate sites to host a landmark) as $\mathcal{A}$, the demographic landmark placement problem can now be formally defined. The problem is to find a set of landmarks $\mathcal{L} \subseteq \mathcal{A}$ having K landmarks subject to an optimization condition. We consider two different approaches to determine the placement of the set of landmarks $\mathcal{L}$ of size K taking into account the concentration of users in $\mathcal{A}$. Such approaches have been used to determine the placement of different kinds of facilities like

Table 1: Adopted notation for the maximum covering location problem.

d_{ij}	geographic distance between sites i and j
D_c	coverage distance
h_i	demand (number of users) at agglomeration i
K	number of landmarks to locate
a_{ij}	$\begin{cases} 1 & \text{if candidate site } i \text{ can cover demands at agglomeration } j\text{, that is, if } d_{ij} \leq D_c, \\ 0 & \text{if not.} \end{cases}$
X_j	$\begin{cases} 1 & \text{if we place landmark at candidate agglomeration } j, \\ 0 & \text{if not.} \end{cases}$
Z_i	$\begin{cases} 1 & \text{if agglomeration } i \text{ is covered,} \\ 0 & \text{if not.} \end{cases}$

fire stations, hospitals, or police departments [11]. In the first case, given the maximum distance from an agglomeration to a landmark, one has to determine the minimum number of landmarks needed to satisfy such a constraint. Moreover, if the condition of covering all agglomerations is relaxed, a smaller number of landmarks may cover a large portion of the space of user's concentration as such a concentration is not uniform throughout the agglomerations. Such a model is known as the maximum covering location model [12]. In the second approach, after fixing the number of landmarks (K) to be located, one requires to know the maximum distance from any agglomeration to its nearest landmark. Such a problem is known as the K-center problem [12].

4.1 Maximum Covering Location Model

We consider fixing the coverage distance of each landmark and maximizing the number of covered demands (the hosts to be located) for a limited number of landmarks, resulting in a maximum covering location model. A landmark covering a certain region is a location estimation for the nodes within that region. The problem of placing landmarks is modeled to take into account the concentration of users within the candidate sites. Table 1 presents the notation adopted to model the problem. Using this notation, the maximum covering location model is formulated by the objective function

$$\max \sum_i h_i Z_i. \tag{3}$$

The objective function (3) maximizes the number of covered demands, that is, the number of users nearby a placed landmark. The objective function is subject to the following constraints:

$$Z_i \leq \sum_j a_{ij} X_j \qquad \forall i \tag{4}$$

$$\sum_j X_j \leq K \tag{5}$$

$$X_j = 0, 1 \qquad \forall j \tag{6}$$

$$Z_i = 0, 1 \qquad \forall i \tag{7}$$

Constraint (4) states that users at agglomeration i are covered if at least one site that covers agglomeration i is selected to host a landmark. Constraint (5) stipulates that we locate no more than K landmarks. Constraints (6) and (7) are the integrality constraints on the decision variables X and Z. We adopt the greedy approach outlined in Algorithm 1 to solve the problem. After assigning the agglomeration of the first landmark to cover the most uncovered demand, the algorithm greedily looks for the best location for the next landmark until K landmarks are placed.

Algorithm 1 Greedy approach to the maximum covering location model [12]

$\mathcal{L} \leftarrow \emptyset$; $\mathcal{A}' \leftarrow \mathcal{A}$
while ($|\mathcal{L}| < K$) **do**
    Find agglomeration $A \in \mathcal{A}'$ that covers the most uncovered demand
    $\mathcal{L} \leftarrow \mathcal{L} \cup A$
    Set $C \subseteq \mathcal{A}'$ to be the set of all agglomerations covered by A
    $\mathcal{A}' \leftarrow \mathcal{A}' - C$
end while
$\mathcal{L}$ is the set of K landmarks

Figure 3 presents the number of landmarks needed to cover a certain percentage of the considered space of users for different values of coverage distance. The incremental user coverage decreases as additional landmarks are assigned. There is a tradeoff between the number of landmarks and the accuracy of the location information represented by the coverage distance. For larger coverage distances, one landmark acts as a location estimation for a larger amount of users, but at a lower accuracy. As a consequence, depending on the requirements on location accuracy of the demanding application, a limited number of landmarks may be enough to provide the desired information.

Figure 4 compares the coverage of users achieved by three policies of landmark placement for coverage distances of 10 and 500 Kms: random placement, geographic placement, and demographic placement. Under random placement, the landmarks are randomly selected, disregarding the concentration of users or the proximity between candidate sites. Geographic placement considers no difference among the candidate sites to locate landmarks in terms of user concentration ($h_i = 1, \forall i$). The number of agglomerations within the range given by the coverage distance (D_c) around a candidate site determines the weight of the candidate site. Among candidate sites having equal weight, the elected candidate site to locate the landmark is randomly chosen. Error bars in the results from the random and geographic placement represent the 99% confidence interval. The demographic landmark placement significantly improves the representativeness (in terms of users to be eventually located) of the chosen landmarks when compared with the random or the geographic placement policies.

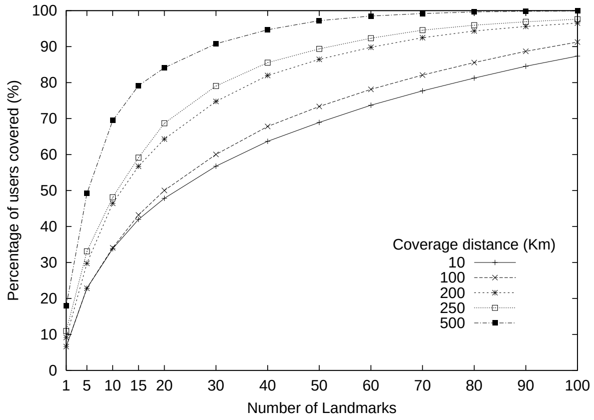
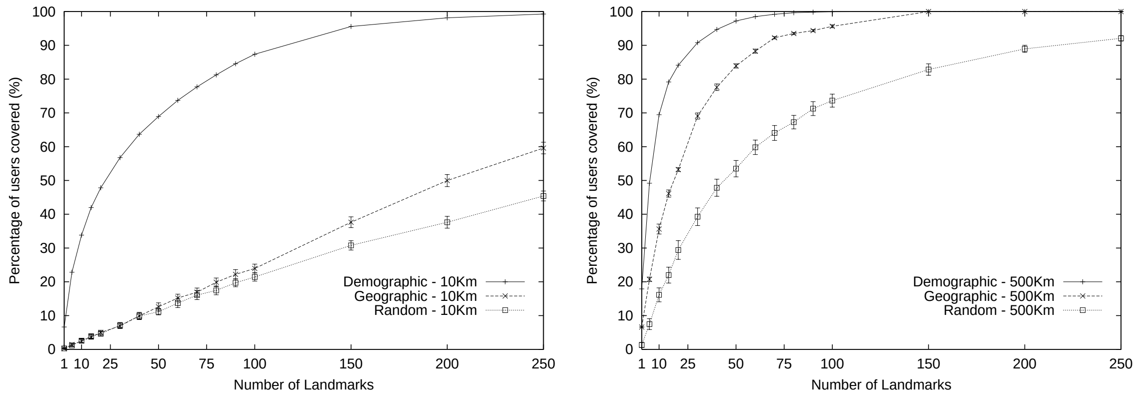


Figure 3: Coverage of users in the maximum coverage location model.



(a) D_c = 10 Km

(b) D_c = 500 Km

Figure 4: Coverage of users in demographic, geographic, and random landmark placement.

The proposed methodology provides not only the number of needed landmarks to cover a certain percentage of users, but the specific geographic locations of such landmarks. To visualize the worldwide geographic location of the landmarks indicated by the demographic placement police, we make use of the GeoPlot tool [13]. Figure 5 illustrates the demographic placement of 50 landmarks to cover almost 90% of the considered space of users for a coverage distance of 250 Km.

We observe in Figure 5 the concentration of landmarks in three main regions: USA, Western Europe, and Japan. The high density of landmarks placed at such regions reflects the concentration of users therein. We can also verify the presence of some areas of relatively high user concentration elsewhere, as in Brazil and Australia. The results indicate the urban agglomerations to place the landmarks to improve the representativeness of each landmark given the constraints on the accuracy and the number of landmarks to be placed.

4.2 K-center Problem

We now consider fixing the number of landmarks that are to be located and minimizing the maximum coverage distance between an agglomeration and the nearest landmark,

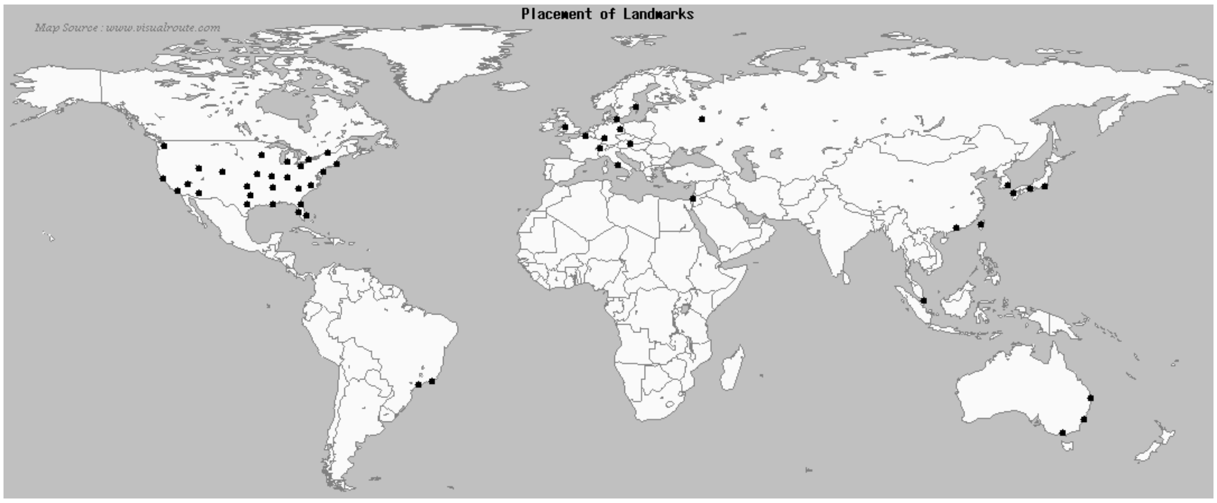


Figure 5: Placement of 50 landmarks for a coverage distance of 250 Km.

Table 2: Additional notation for the K-center problem.

Y_{ij}	$\begin{cases} 1 & \text{if agglomeration } i \text{ is assigned to a landmark at site } j, \\ 0 & \text{if not.} \end{cases}$
W	maximum distance between an agglomeration and the nearest landmark

resulting in a K-center problem. Again, a landmark covering a certain region is a location estimation for the nodes within that region. Table 2 defines some additional notation to model the problem. Using the notation defined in Tables 1 and 2, the K-center problem is formulated by the objective function

$$\min W. \tag{8}$$

The objective function (8) minimizes the maximum distance between an agglomeration and the nearest landmark. The objective function is subject to the following constraints:

$$\sum_j X_j = K \tag{9}$$

$$Y_{ij} \leq X_j \qquad \forall i, j \tag{10}$$

$$W \geq \sum_j d_{ij} Y_{ij} \qquad \forall i \tag{11}$$

$$X_j = 0, 1 \qquad \forall j \tag{12}$$

$$Y_{ij} = 0, 1 \qquad \forall i, j \tag{13}$$

Constraint (9) stipulates that K landmarks are to be located. Constraint (10) states that agglomeration i can only be covered at site j if site j has a landmark. Constraint (11) defines the lower bound on the maximum distance, which is being minimized. In other words, the maximum distance (W) must be greater or equal than the distance between any agglomeration i and the landmark j to which the agglomeration i is assigned. Constraint (12) is the integrality constraint on the decision variable X. Constraint (13) requires the agglomeration i to be assigned to only one landmark j.

Even if an agglomeration i is within the coverage distance of a landmark j ($a_{ij} = 1$), the agglomeration i may be assigned to another closer landmark k, resulting in $Y_{ij} = 0$. Therefore, we notice in particular that $a_{ij} \geq Y_{ij}$. This is the case as there is no reason to assign an agglomeration to a landmark other than the closest one. Considering unweighted distances leads to minimize the geographic distance between an agglomeration and the nearest landmark no matter the concentration of users are and is equivalent to a geographic placement of landmarks.

In the demographic landmark placement, we want to consider the demand-weighted distances (taking into account the concentration of users at each agglomeration) between agglomerations and candidate sites to host landmarks. In such a case, constraint (11) should be replaced by

$$W \geq h_i \sum_j d_{ij} Y_{ij}, \qquad \forall i. \tag{14}$$

One solution to the problem is to enumerate each possible subset of size K out of the $|\mathcal{A}|$ candidate locations and then verify which provides the minimum value of W. Nevertheless, even for moderate values of $|\mathcal{A}|$ and K such enumeration is not realistic, as the K-center problem is known to be NP-Complete [14]. An alternative binary search approach to solve the K-center problem is outlined by Algorithm 2. The approach is based on searching the smallest coverage distance that allows all agglomerations to be covered.

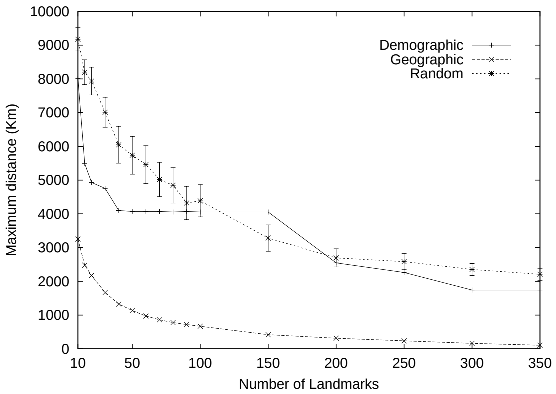
Algorithm 2 Binary search approach to the K-center problem [12]

1: $D_c^L \leftarrow 0$
2: $D_c^H \leftarrow \max_{i,j}\{d_{ij}\}$
3: **while** ($D_c^H \neq D_c^L$) **do**
4: $\quad D_c = \lfloor (D_c^H + D_c^L)/2 \rfloor$
5: $\quad$ Compute the smallest $\mathcal{L}(D_c)$ set of landmarks that covers the entire set of agglomerations $\mathcal{A}$ for a coverage distance of D_c.
6: $\quad$ **if** ($|\mathcal{L}(D_c)| \leq K$) **then**
7: $\qquad D_c^H \leftarrow D_c$
8: $\quad$ **else**
9: $\qquad D_c^L \leftarrow D_c + 1$
10: $\quad$ **end if**
11: **end while**
12: D_c^L is the solution to the objective function and $\mathcal{L}(D_c^L)$ provides the locations of landmarks for the solution.

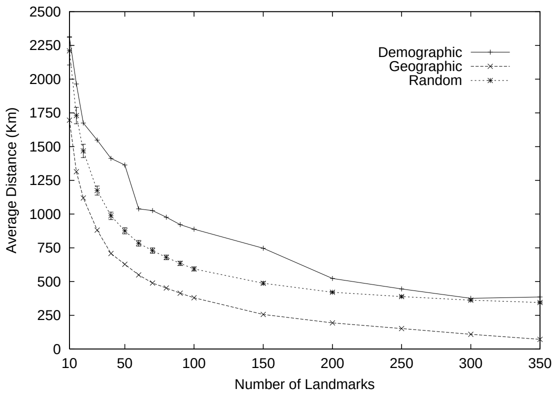
Algorithm 2 considers unweighted distances resulting in a geographic landmark placement. To consider demand-weighted distances as in the case of the demographic landmark placement, the initial value of D_c^H needs to be set to take into account the weighted distance. Therefore, step 2 in Algorithm 2 is replaced by

$$D_c^H \leftarrow \left[\max_{i,j}\{d_{ij}\}\right]\left[\max_i\{h_i\}\right]$$

and in finding the smallest set of landmarks that cover the entire set of agglomerations (step 5 in Algorithm 2), a candidate site to host a landmark j will be able to cover an agglomeration i if $d_{ij} h_i \leq D_c$.



(a) Maximum distance



(b) Average distance

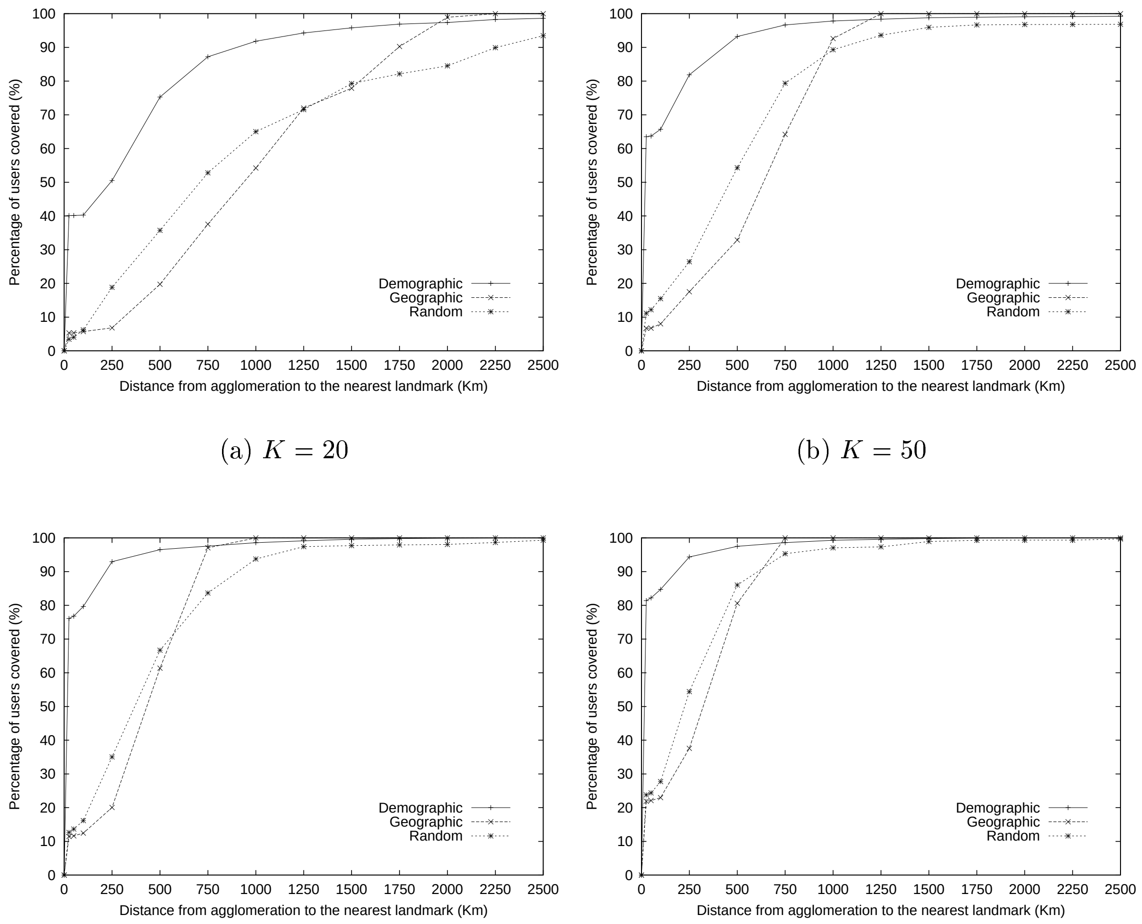
Figure 6: Distance from an agglomeration to the nearest landmark (Km).

Figure 6 presents the maximum distances W and average distances between an agglomeration and its nearest landmark for a different number of landmarks. The presented maximum distances reflect the worst case distance for the whole set of agglomerations. The average distances between each agglomeration hosting no landmark and the correspondent nearest landmark are also depicted in Figure 6(b).

We compare the results from the demographic landmark placement, the geographic placement, and the random placement policies. The geographic placement performs the best providing the lowest maximum distances between agglomerations and the nearest landmark. The demographic landmark placement provides higher maximum and average distances, but such results hide the concentration of users in the agglomerations with such assigned distances. Adding more landmarks may even not change the maximum distance observed under demographic landmark placement. In such a case, a remote agglomeration with low user concentration is kept far from the nearest landmark as additional landmarks are used to decrease the distance between more user populated agglomerations to the nearest landmark. We observe such situation in Figure 6(a) between 50 and 150 placed landmarks as the maximum distance keeps leveled off. For instance, the remote low user density agglomeration is located in Nigeria, Africa. However, the average distance presented in Figure 6(b) keeps decreasing in the same range of additional landmarks as the density of landmarks in denser user areas increases.

The demographic landmark placement adopts weighted distances to consider the user concentrations within the different agglomerations. Such a procedure pushes the worst case distances towards the farthest and least user concentrated agglomerations. High user concentrated agglomerations get closer landmarks and the agglomerations having the highest concentrations host the landmarks. The demographic landmark placement provides smaller distances between the agglomerations and the nearest landmark for the most part of users at the expense of leaving farther landmarks for the remote and least user concentrated agglomerations. Such results are depicted in Figure 7 for a different number of placed landmarks.

Figure 8 illustrates the difference in adopting either unweighted (geographic placement) or weighted (demographic landmark placement) distances to be minimized in the K-center problem. Figure 8(a) presents the placement of 50 landmarks considering ge-



(a) $K = 20$

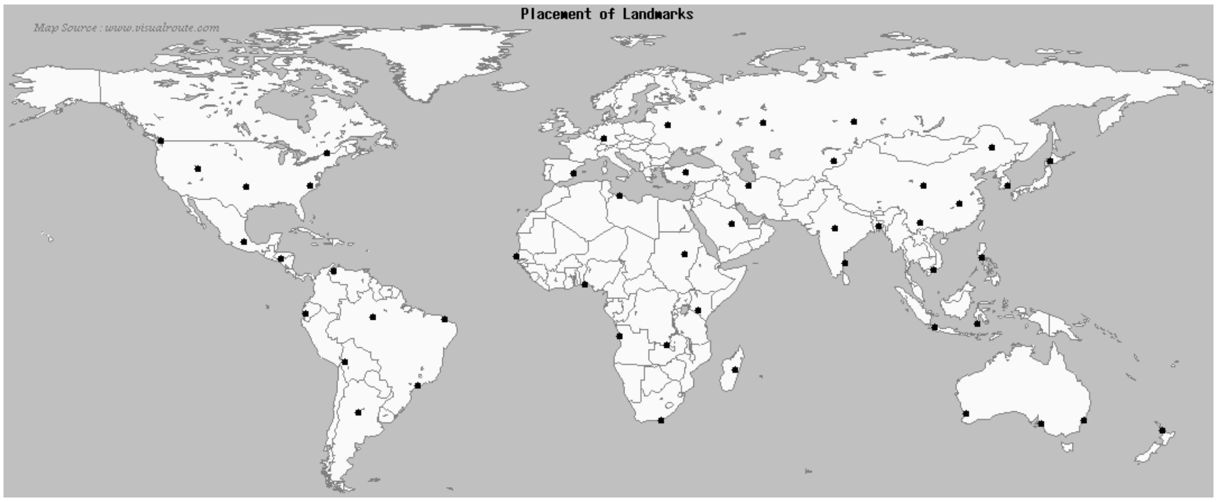
(b) $K = 50$

(c) $K = 80$

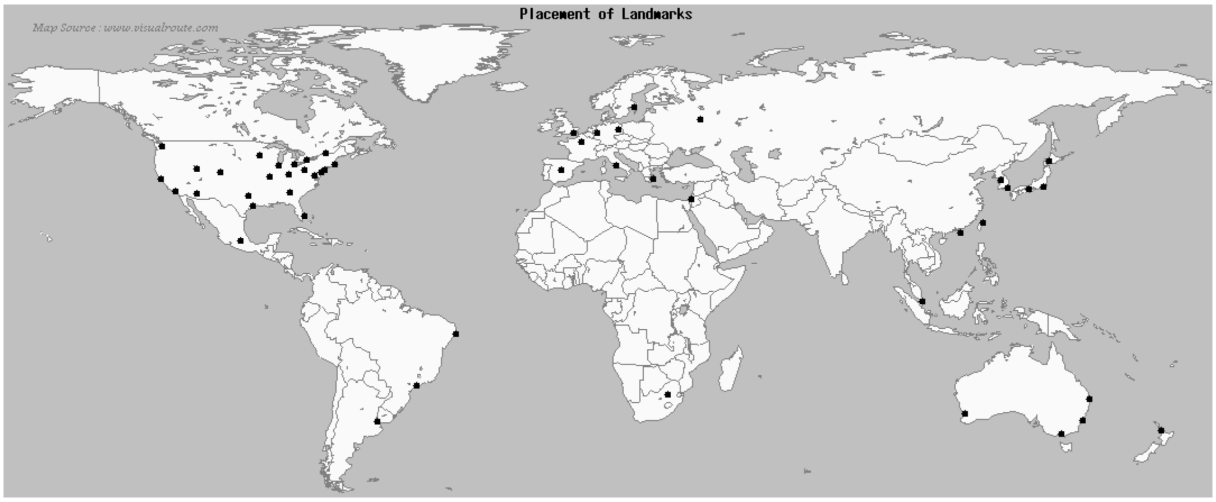
(d) $K = 100$

Figure 7: Percentage of covered users as a function of the distance to the nearest landmark.

ographic landmark placement, that is, disregarding the concentration of users. Such a disregard of user's concentration leads to a more uniform geographic distribution of the 50 landmarks worldwide and to a lower maximum (unweighted or purely geographic) distance W between any agglomeration and its nearest landmark. In the other hand, Figure 8(b) presents the placement of 50 landmarks considering weighted distances as a consequence of the demographic landmark placement. The maximum (unweighted or purely geographic) distance W between an agglomeration and the nearest landmark is higher in this case, but the worst cases are pushed to the remote agglomerations having the lowest concentrations of users. The result is a dense placement of landmarks in the regions presenting the highest concentrations of users like the USA, Western Europe, and Japan.



(a) Geographic landmark placement (disregards concentration of users)



(b) Demographic landmark placement (considers concentration of users)

Figure 8: K-center solution for placing 50 landmarks.

5 Related Work and Further Applicability

Techniques for geographically locating an Internet host are studied in [3]. Such techniques include the inference from DNS names, from clustering the IP address space, and from measurements. Nevertheless, the study about inference from measurements does not consider the issue of landmark placement and is basically US-centric. Recent works focus on where to place Internet resources other than landmarks like caches [15] and mirrors [16]. In [8], authors investigate the geographic location of Internet components to create a base for the development of geographically-driven topology generation methods.

Frameworks for estimating distance between two Internet hosts have been recently proposed, including the IDMaps [17] proposal and the Global Network Positioning (GNP) scheme [18]. Nevertheless, such structures provide distance between two hosts in terms of delay, not geographic distance. As discussed in Section 2, the correlation between delay and geographic distance is weak worldwide. Therefore, even if such a correlation tends to be stronger within regions having richer connectivity, the distance estimations in terms of delay provided by IDMaps or GNP can not be directly applied to infer geographic distances. However, structures like IDMaps or GNP may be used to feed the location inference from measurements with delay estimations between the appropriate hosts (probe machines, landmarks, and host to be located). More accurate delay measurements provided by IDMaps or GNP may result in better data for techniques based on location inference from measurements.

Topologically-aware construction of overlay networks and server selection [19] adopts landmarks as references of proximity to group nodes that are relatively close to one another in terms of network delay. Such binning strategy avoids unnecessary high latency hops, thus improving routing performance. We expect that demographically placing the reference landmarks closer to where most users are, as proposed in this study, would improve the performance of the binning strategy as more users would be closer to the reference landmarks.

6 Conclusion

In this paper, we have discussed the problem of landmark placement for Internet host location. The demographic landmark placement has been proposed to handle the location problem taking into account where users (hosts to be located) are. Considering the demographic placement, we have assessed two different approaches to tackle the location problem. In the former one, we have considered fixing the coverage distance of each landmark and maximizing the number of potential hosts to be located for a limited number of landmarks. In the latter one, we have considered fixing the number of landmarks to be placed and minimizing the maximum coverage distance between a user agglomeration and the nearest landmark.

Results show that the demographic landmark placement allows a relatively small number of landmarks to represent a large portion of users within a limited coverage distance. Fewer landmarks also imply a lower amount of measurement traffic in the network. There is a diminishing return in the coverage of users for each extra landmark. For a fixed number of landmarks, the demographic landmark placement enhances the representativeness of each landmark. Such landmarks are placed on areas of high user density, thus providing closer landmarks and more accurate location estimations for the most part of hosts to be located. As future work, we intend to investigate hierarchical approaches to mitigate the impact of the measurement traffic on the network.

References

[1] M. Weiser, "The computer for the 21st century", *Scientific American*, vol. 265, pp. 94–104, Sept. 1991.

[2] C. Davis, P. Vixie, T. Goowin, and I. Dicknson, "A means for expressing location information in the domain name system", *Internet RFC 1876*, Jan. 1996.

[3] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts", in *Proc. of the ACM SIGCOMM'2001*, (San Diego, CA, USA), Aug. 2001.

[4] D. Moore, R. Periakaruppan, J. Donohoe, and K. Claffy, "Where in the world is netgeo.caida.org?", in *Proc. of the INET'2000*, (Yokohama, Japan), July 2000.

[5] *Library Servers via WWW – LibWeb.* `http://sunsite.berkeley.edu/Libweb`.

[6] L. Subramanian, V. N. Padmanabhan, and R. Katz, "Geographic properties of Internet routing: Analysis and implications", in *Proc. of USENIX 2002*, (Monterey, CA, USA), June 2002.

[7] S.-H. Yook, H. Jeong, and A.-L. Barabási, "Modeling the Internet's large-scale topology", *Proceedings of the National Academy of Science (PNAS)*, 2002. accepted for publication.

[8] A. Lakhina, J. W. Byers, M. Crovella, and I. Matta, "On the geographic location of Internet resources", in *ACM Internet Measurement Workshop 2002*, (Marseille, France), Nov. 2002.

[9] T. Brinkhoff, *City Population.* `http://www.citypopulation.de`.

[10] Central Intelligence Agency (CIA), *The World Factbook 2001*, Jan. 2001. `http://www.cia.gov/cia/publications/factbook`.

[11] C. Toregas, R. Swain, C. Revelle, and L. Bergman, "The location of emergency service facilities", *Operations Research*, vol. 19, no. 6, pp. 1363–1373, 1971.

[12] M. S. Daskin, *Network and Discrete Location.* New York, NY: John Wiley & Sons, 1995.

[13] CAIDA, *GeoPlot.* `http://www.caida.org/tools/visualization/geoplot/`.

[14] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness.* New York, NY: W. H. Freeman & Co., 1979.

[15] P. Krishnan, D. Raz, and Y. Shavitt, "The cache location problem", *IEEE/ACM Transactions on Networking*, vol. 8, pp. 568–582, Oct. 2000.

[16] E. Cronin, S. Jamin, C. Jin, A. R. Kurc, D. Raz, and Y. Shavitt, "Constrained mirror placement on the Internet", *IEEE Journal on Selected Areas in Communications*, vol. 20, pp. 1369–1382, Sept. 2002.

[17] P. Francis, S. Jamin, C. Jin, Y. Jin, D. Raz, Y. Shavitt, and L. Zhang, "IDMaps: A global Internet host distance estimation service", *IEEE/ACM Transactions on Networking*, vol. 9, pp. 525–540, Oct. 2001.

[18] T. S. E. Ng and H. Zhang, "Predicting Internet network distance with coordinates-based approaches", in *Proc. of the IEEE INFOCOM'2002*, (New York, NY, USA), June 2002.

[19] S. Ratnasamy, M. Handley, R. Karp, and S. Shenker, "Topologically-aware overlay construction and server selection", in *Proc. of the IEEE INFOCOM'2002*, (New York, NY, USA), June 2002.