Demographic Placement for Internet Host Location

Artur Ziviani*†, Serge Fdida*
*Laboratoire d'Informatique de Paris 6 (LIP6)
Université Pierre et Marie Curie
Paris, France
Email: {Artur.Ziviani, Serge.Fdida}@lip6.fr

José F. de Rezende†, Otto Carlos M. B. Duarte†
†Grupo de Teleinformática e Automação (GTA)
COPPE/EE – Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
Email: {rezende, otto}@gta.ufrj.br

***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 infers host locations using delay measurements to geographically distributed landmarks, which are hosts with a known geographic location. The problem we deal with is where to place such landmarks and the probe machines that perform the delay measurements. We propose a demographic placement approach to improve the representativeness of each landmark with respect to the hosts to be located. Results show that a relatively small number of landmarks are sufficient to cover the most part of hosts to be located. For a fixed number of landmarks, the demographic approach reduces the distances from most hosts to the nearest landmark. Considering the probe machines, we show that they have to be sparsely placed to avoid gathering redundant data.**

I. Introduction

Knowing the geographic location of an Internet host from an identifier of that host, such as a name or IP address, enables a whole new class of location-aware applications. Examples of these new Internet applications are targeted advertising on web pages, automatic selection of a language to first display the content, or authorization of transactions only when performed from pre-established locations. Different techniques [1] infer the location of an Internet host from DNS names, from clustering the IP address space, or from delay measurements.

The location estimation of a host from delay measurements is based on the observation that hosts sharing similar delays to some fixed probe machines tend to be near each other geographically. Thus, given a set of landmarks, which are hosts with a known geographic location, the location estimation for a target host is the location of the landmark with the most similar delay pattern to the target host. Given a finite number of landmarks, we are interested in where to place them to provide better location estimations for the majority of users (hosts).

We propose a demographic placement approach that considers the geographic distribution of users, and consequently of hosts to be located, to place landmarks. We also deal with the placement of probe machines that perform delay measurements to landmarks and hosts to be located.

Results show that the demographic placement provides a relatively small number of landmarks able to represent a large portion of users within a limited coverage distance. As a result, fewer landmarks imply a lower amount of measurement traffic. For a fixed number of landmarks, the demographic placement improves the representativeness of each landmark, resulting in closer landmarks and more accurate location estimations. Probe machines are placed on sites likely to have sufficient network infrastructure to make their deployment feasible and to avoid gathering redundant data.

This paper is organized as follows. In Section II, we formalize the problem of host location inference from delay measurements. Section III presents and evaluates the demographic placement proposition. The related work is discussed in Section IV. In Section V, we present our conclusions.

II. Inferring Host Locations from Measurements

Inferring a host geographic location from delay measurements was first introduced by GeoPing [1] and can be formalized as follows. Consider a set $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ of K landmarks, which are hosts with a known geographic location. Also consider a set $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$ of N probe machines. The probe machines periodically determine the delay, which is actually the minimum delay from several measurements, to each landmark. Therefore, each probe machine $P_x \in \mathcal{P}$ keeps a delay vector $\mathbf{d}_x = (d_{1x}, d_{2x}, \ldots, d_{Kx})$.

Suppose one desires to determine the location of a given target host T. A location server that knows the set of landmarks $\mathcal{L}$ and of probe machines $\mathcal{P}$ is then contacted to ask the N probe machines to measure the delay to host T. Each probe machine $P_x \in \mathcal{P}$ returns a delay vector $\mathbf{d}'_x = (d_{1x}, d_{2x}, \ldots, d_{Kx}, d_{Tx})$, i.e., the delay vector $\mathbf{d}_x$ plus the just measured delay to host T. The location server is then able to construct the delay matrix $\mathbf{D}$ with dimensions $(K+1) \times N$:

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

The delay vectors gathered by the location server from the N probe machines correspond to the columns of the delay matrix $\mathbf{D}$. The location server then compares the lines of the delay matrix $\mathbf{D}$ to estimate the location of host T. To infer the location of host T, the nearest landmark with respect to T is determined using Euclidean distance. The landmark L having the smallest Euclidean distance $e_{LT} = \sqrt{(d_{y1} - d_{T1})^2 + (d_{y2} - d_{T2})^2 + \ldots + (d_{yN} - d_{TN})^2}$ from host T, where $y = 1, \ldots, K$, is the nearest landmark with respect to T. The corresponding location of the landmark L is the location estimation of host T. The accuracy of the location estimation and the network load due to measurements basically depend on the location and on the number of landmarks. Therefore, we are interested in where to place a finite number of landmarks in order to reduce measurement traffic load while increasing the representativeness of each placed landmark.

III. Demographic Placement

A landmark is a reference to be used as a location estimation of a certain number of hosts supposed to be nearly located. Thus, landmarks located in areas with high density of hosts provide location estimations that reflect the positions of a large number of hosts. Landmarks can be any oblivious host able to echo `ping` messages and with a known location. In the demographic approach, we place the landmarks according to the user (host) population distribution. Recent findings [2], [3] indicate a strong correlation between population and router density in economically developed countries.

We consider the 407 main urban agglomerations spread worldwide [4] since they are likely to offer the highest concentration of users (hosts to be located). Internet infrastructure varies dramatically across different regions throughout the world. Therefore, we weight the populations of the different agglomerations with the number of users in the country the agglomeration belongs to over the total population of the country. In applying this weight, we estimate the 407 main user agglomerations worldwide and their demands to be covered by the demographic strategy. Data on the user and total population of each country are available in [5].

We denote as $\mathcal{A}$ the set of user agglomerations, i.e. the candidate sites to host a landmark. The set $\mathcal{A}$ represents the 407 main user agglomerations, totalizing 173,696,253 estimated users [4], [5]. The mean distance between each pair of elements in $\mathcal{A}$ is 8167 km. The demographic landmark placement problem is then to find a set of landmarks $\mathcal{L} \subseteq \mathcal{A}$ with K landmarks subject to an optimization condition. We investigate two complementary approaches to place the set of landmarks $\mathcal{L}$ of size K taking into account the concentration of users in $\mathcal{A}$. Table I presents the adopted notation to model the problem of placing landmarks and probe machines.

A. Maximum Covering Location Model (MCLM)

We first 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 (MCLM) [6]. The demographic placement of landmarks takes into account the concentration of users within the candidate sites. Using the notation from Table I, the maximum covering location model is formulated by the objective function

$$\max \sum_i h_i Z_i. \tag{2}$$

The objective function (2) maximizes the number of covered demands, i.e. the number of users nearby a placed landmark, and it is subject to the following constraints:

TABLE I

Adopted notation.

g_{ij}	geographic distance between agglomerations i and j
h_i	demand (number of users) at agglomeration i
G	geographic coverage distance
W	maximum distance from any agglomeration to a landmark
M	minimum distance between any pair of placed probe machines
a_{ij}	$\begin{cases} 1 & \text{if candidate site } i \text{ can cover demands at agglomeration } j, \\ 0 & \text{if not.} \end{cases}$
X_i	$\begin{cases} 1 & \text{if one places a landmark on candidate site } i, \\ 0 & \text{if not.} \end{cases}$
Z_i	$\begin{cases} 1 & \text{if agglomeration } i \text{ is covered}, \\ 0 & \text{if not.} \end{cases}$
Y_{ij}	$\begin{cases} 1 & \text{if agglomeration } i \text{ is assigned to a landmark at site } j, \\ 0 & \text{if not.} \end{cases}$
Q_i	$\begin{cases} 1 & \text{if one places a probe machine on candidate site } i, \\ 0 & \text{if not.} \end{cases}$

Algorithm 1 Greedy approach to the MCLM [6]

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

$$Z_i \leq \sum_j a_{ij} X_j \qquad \forall i \tag{3}$$

$$\sum_j X_j \leq K \tag{4}$$

We adopt the greedy approach outlined in Algorithm 1 to solve the MCLM problem with time complexity $O(|\mathcal{A}|^2 K)$. The algorithm greedily places landmarks to cover the most uncovered demand until K landmarks are placed.

Fig. 1 presents the number of landmarks needed to cover a certain percentage 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, which is represented by the coverage distance. For large coverage distances, the landmark location acts as an estimation for a large 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 is enough to provide the desired information.

For a coverage distance of 10 km, Fig. 2 compares the coverage of users achieved by three policies of landmark placement: random, geographic, and demographic. Under random placement, the landmarks are randomly selected, disregarding the concentration of users or the proximity between candidate sites. Geographic placement considers no difference in terms of user concentration ($h_i = 1, \forall i$) among candidate sites to host landmarks. The number of agglomerations within the range given by the coverage distance G around one candidate site determines its weight. Among candidate sites with equal weight, the elected site is randomly chosen. Error bars in the results from the random and geographic placement represent

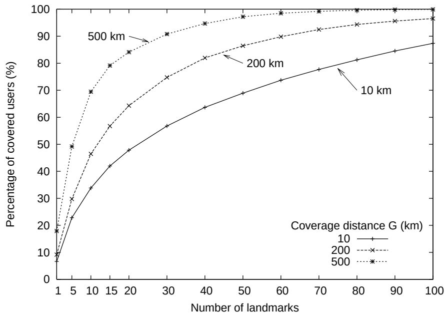


Fig. 1. Covered users in the demographic landmark placement.

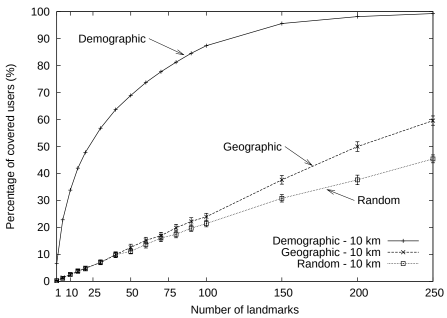


Fig. 2. Covered users in demographic, geographic, and random placements.

the 99% confidence interval. The demographic scheme significantly improves the representativeness in hosts to be eventually located of each placed landmark.

B. K-center Problem

We now tackle the problem of minimizing the maximum distance between an agglomeration and the nearest landmark while considering a fixed number of landmarks. One can notice that even if an agglomeration i is within the coverage distance of a landmark j ($a_{ij} = 1$), it may be assigned to another closer landmark k ($Y_{ij} = 0$ and $Y_{ik} = 1$). Therefore, $a_{ij} \geq Y_{ij}$ as there is no reason to assign an agglomeration to a landmark other than the closest one. This minimization problem is known as the K-center problem [6], which can be formulated using the notation defined in Table I by the objective function

$$\min W \tag{5}$$

and it is subject to the following constraints:

$$\sum_j X_j = K \tag{6}$$

$$Y_{ij} \leq X_j \qquad \forall i, j \tag{7}$$

$$W \geq \sum_j g_{ij} Y_{ij} \qquad \forall i \tag{8}$$

In the demographic approach, in order to consider the concentration of users in each agglomeration, the constraint (8) should be replaced by $W \geq h_i \sum_j g_{ij} Y_{ij}, \forall i$. Otherwise, the minimization problem leads to a geographic placement of landmarks.

Algorithm 2 Binary search approach to K-center problem [6]

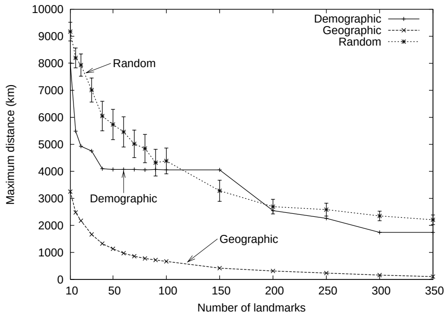
1: $G_{\max} \leftarrow \max_{i,j}\{g_{ij}\}$
2: $G_L \leftarrow 0$; $G_H \leftarrow G_{\max}$
3: **while** $(G_H \neq G_L)$ **do**
4: $\quad G \leftarrow \lfloor (G_H + G_L)/2 \rfloor$
5: $\quad$ Find the smallest set of landmarks $\mathcal{L}(G)$ that covers $\mathcal{A}$ for a coverage distance of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

The K-center problem is known to be NP-Complete [7]. An alternative binary search approach is outlined in Algorithm 2 to solve the K-center problem with time complexity $O(|\mathcal{A}|K^2 \log G_{\max})$. Such an approach defines a lower bound G_L and an upper bound G_H on the maximum distance between an agglomeration and the nearest landmark. The approach then successively narrows the range between such bounds until they converge into the smallest coverage distance that allows a set of K landmarks to cover all agglomerations. This coverage distance is the smallest maximum distance between an agglomeration and the nearest landmark for K placed landmarks, thus being the solution of the objective function (5). The resulting set of K landmarks is the approximate solution to the K-center problem.

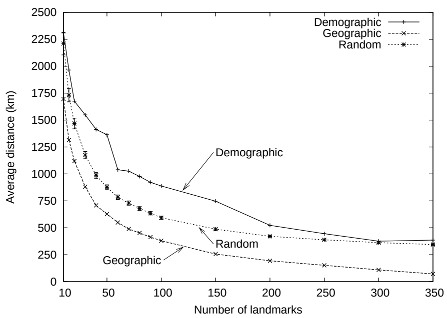
Algorithm 2 considers unweighted distances, resulting in a geographic placement of landmarks. In the demographic strategy, to consider demand-weighted distances, step 1 in Algorithm 2 should state $G_{\max} \leftarrow \Big[\max_{i,j}\{g_{ij}\}\Big]\Big[\max_i\{h_i\}\Big]$. Furthermore, in computing the smallest set of landmarks that covers the entire set of agglomerations (step 5 in Algorithm 2), a candidate site to host a landmark j is able to cover an agglomeration i if $g_{ij}h_i \leq G$.

We compare the results from the random, geographic, and demographic placement policies in Fig. 3. The geographic placement performs the best, providing the lowest maximum geographic distances between agglomerations and the nearest landmark. The demographic approach provides higher maximum and average distances, but such results mask the concentration of users in the agglomerations. Adding more landmarks may even not change the maximum distance observed under the demographic placement policy. A remote agglomeration with low user concentration may be kept far from the nearest landmark as additional landmarks are used to decrease the distance between more user populated agglomerations to the nearest landmark. One observes this situation in Fig. 3(a) between 50 and 150 placed landmarks as the maximum distance keeps leveled off. Nevertheless, as shown in Fig. 3(b), the average distance to the nearest landmark keeps decreasing in this same range as the density of landmarks in denser user areas increases.

The demographic placement considers the user concentrations within the different agglomerations. As a consequence, the demographic placement pushes the worst-case distances



(a) Maximum geographic distance



(b) Average geographic distance

Fig. 3. Distance from an agglomeration to the nearest landmark (km).

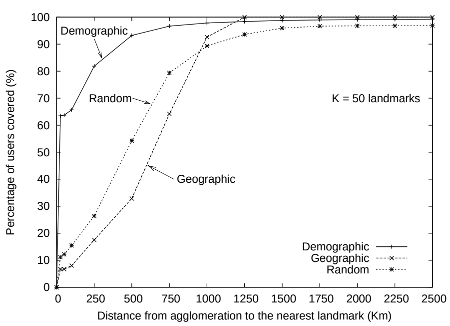


Fig. 4. Covered users as a function of the distance to the nearest landmark.

toward the farthest and least user concentrated agglomerations. High user concentrated agglomerations are assigned to closer landmarks and the agglomerations with the highest user concentrations host the landmarks. Such results are presented in Fig. 4 using 50 placed landmarks. The demographic strategy provides smaller distances between the agglomerations and their nearest landmark for the most part of users at the expense of leaving farther landmarks for the remote and least user concentrated agglomerations.

C. Placement of Probe Machines

In this subsection we deal with the issue of how and where to place the probe machines. Probe machines measure the delay to a target host to be located and regularly gather delays to landmarks. Probe machines that are near each other may share

Algorithm 3 Greedy approach to place probe machines

$\mathcal{A}' \leftarrow \mathcal{A}$; $\mathcal{P} \leftarrow A \in \mathcal{A}$ with $\max_i\{h_i\}$
while $(|\mathcal{P}| < N)$ **do**
  Find $A \in \mathcal{A}'$ that has the maximum (weighted) distances toward the previously placed probe machines.
  $\mathcal{P} \leftarrow \mathcal{P} \cup A$; $\mathcal{A}' \leftarrow \mathcal{A}' - A$
end while
$\mathcal{P}$ is the set of N probe machines

common paths to some remote landmarks or target hosts, thus providing redundant information to the location estimation decision. Therefore, our first goal is to make a geographically sparse distribution of probe machines to avoid shared paths. In order to make the deployment of probe machines feasible, they are likely to be placed on agglomerations with better network infrastructure. The second goal is thus to maximize the number of users on the agglomerations selected to host probe machines. We adopt the number of users (hosts) as a means to reflect the level of network infrastructure on a given agglomeration.

The problem of placing probe machines is to find a set $\mathcal{P} \subseteq \mathcal{A}$ with N probe machines that maximizes the minimum weighted distance between any pair of placed probe machines. We adopt the weighted distance between agglomerations to consider the user concentration at each agglomeration and thus place probe machines on agglomerations likely to have better network infrastructure. Using the notation defined in Table I, the problem of placing probe machines is formulated by the objective function

$$\max M. \tag{9}$$

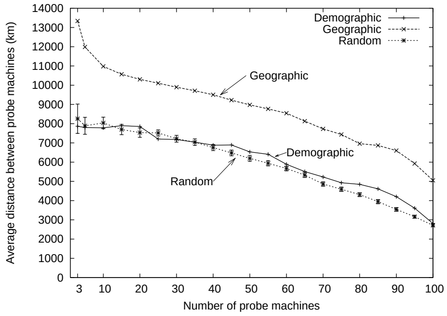
The objective function (9) maximizes the minimum distance between any pair of placed probe machines and it is subject to the following constraints:

$$\sum_i Q_i = N \tag{10}$$

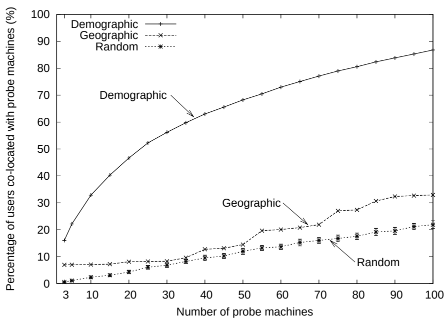
$$M \leq h_i \sum_j g_{ij} Q_j \qquad \text{if } Q_i = 1, \forall i \tag{11}$$

We propose the greedy approach outlined in Algorithm 3 to solve the problem of placing probe machines with time complexity $O(|\mathcal{A}|^2 N)$. The set of probe machines is initialized with the agglomeration presenting the highest user concentration. Afterwards the algorithm greedily places a new probe machine to maximize the weighted distance between the new probe machine and the previously placed probe machines until N are placed. Under the geographic placement, in order to consider unweighted distances, constraint (11) is replaced by $M \leq \sum_j g_{ij} Q_j$, if $Q_i = 1, \forall i$.

Fig. 5 presents the results to the problem of placing probe machines for the demographic, geographic, and random placement policies. In Fig. 5(a), we compare the average distance observed between all pairs of the N placed probe machines. The geographic placement performs the best in sparsely distributing the probe machines as it presents the largest average distances. Nevertheless, such results are obtained placing probe machines on remote user agglomerations



(a) Average distance between probe machines



(b) Users co-located with probe machines

Fig. 5. Placement of probe machines.

that not necessarily have a good network infrastructure. In Fig. 5(b), one observes that the demographic approach places the probe machines on agglomerations with high density of users. Such agglomerations are likely to have a sufficient network infrastructure to make the deployment of probe machines feasible. This result is obtained while the demographic approach is still able to provide a relatively sparse distribution of probe machines in order to avoid the shared paths toward landmarks and hosts to be located as shown in Fig. 5(a).

IV. Related Work

Techniques for geographically locating an Internet host are studied in [1]. These techniques include the inference from DNS names, from clustering the IP address space, and from delay measurements. Nevertheless, the study about location inference from delay measurements disregards the issue of placing landmarks and probe machines. Recent works focus on where to place Internet resources other than landmarks like caches [8] or mirrors [9]. We study similarity models for Internet host location in [10]. The bases for topology generation methods that take into account the geographic location of Internet resources are investigated in [3]. RFC 1876 [11] proposes to add host locations to the DNS records. This proposition, however, is not widely adopted since it requires changes in DNS structure and administrators have no motivation to register new location records.

Frameworks for estimating distance between two Internet hosts have been recently proposed, such as IDMaps [12] and Global Network Positioning (GNP) [13]. Nevertheless, these frameworks provide distance between two hosts in terms of delay, not geographic distance. In spite of that, more accurate delay estimations between the appropriate hosts (probe machines, landmarks, and host to be located) may be used to provide better data for techniques based on location inference from delay measurements.

V. Conclusion

In this paper, we have dealt with the problem of placing landmarks and probe machines for Internet host location. Results show that the proposed demographic 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. 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. Probe machines are placed on locations likely to have sufficient network infrastructure and sparsely distributed to avoid gathering redundant data.

Acknowledgment

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

References

[1] 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.

[2] S.-H. Yook, H. Jeong, and A.-L. Barabási, "Modeling the Internet's large-scale topology," *Proc. of the National Academy of Sciences (PNAS)*, vol. 99, pp. 13 382–13 386, Oct. 2002.

[3] A. Lakhina, J. W. Byers, M. Crovella, and I. Matta, "On the geographic location of Internet resources," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 6, pp. 934–948, Aug. 2003.

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

[5] *The World Factbook 2002*, Central Intelligence Agency (CIA), Jan. 2002, http://www.cia.gov/cia/publications/factbook.

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

[7] 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.

[8] P. Krishnan, D. Raz, and Y. Shavitt, "The cache location problem," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 568–582, Oct. 2000.

[9] 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, no. 7, pp. 1369–1382, Sept. 2002.

[10] A. Ziviani, S. Fdida, J. F. de Rezende, and O. C. M. B. Duarte, "Similarity models for Internet host location," in *Proc. of the IEEE International Conference on Networks - ICON'2003*, Sydney, Australia, Sept. 2003.

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

[12] 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, no. 5, pp. 525–540, Oct. 2001.

[13] 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.