Fundamental Effects of Clustering on the Euclidean Embedding of Internet Hosts *

Sanghwan Lee[1], Zhi-Li Zhang[2], Sambit Sahu[3], Debanjan Saha[3], and Mukund Srinivasan[2]

[1] Kookmin University, Seoul, Korea sanghwan@kookmin.ac.kr
[2] University of Minnesota, Minneapolis, MN, USA {zhzhang,mukund}@cs.umn.edu
[3] IBM T.J. Watson Research Center, Hawthorne, NY, USA
{sambits,dsaha}@us.ibm.com

Abstract. The network distance estimation schemes based on Euclidean embedding have been shown to provide reasonably good overall accuracy. While some recent studies have revealed that *triangle inequality violations* (TIVs) inherent in network distances among Internet hosts fundamentally limit their accuracy, these Euclidean embedding methods are nonetheless appealing and useful for many applications due to their simplicity and scalability. In this paper, we investigate why the Euclidean embedding shows reasonable accuracy despite the prevalence of TIVs, focusing in particular on the effect of *clustering* among Internet hosts. Through mathematical analysis and experiments, we demonstrate that clustering of Internet hosts reduces the effective dimension of the distances, hence low-dimension Euclidean embedding suffices to produce reasonable accuracy. Our findings also provide us with good guidelines as to how to select landmarks to improve the accuracy, and explains why random selection of a large number of landmarks improves the accuracy.

1 Introduction

Network distance estimation schemes have been extensively studied during the past several years. These schemes include [1–5], just to name a few. Among many proposed schemes, the coordinate based schemes are gaining interest because of its simplicity and reasonably good accuracy. In the coordinate based system, each host is assigned a set of coordinates. The set of coordinates represents the position of the host in a virtual Euclidean space. The network distance is estimated by the Euclidean distance between the two hosts in the virtual Euclidean space. To assign coordinates to the hosts, many schemes rely on a set of special hosts called landmarks. Each host measures the distances to the landmarks and transforms the measured distances into a set of coordinates by using various optimization techniques.

Although Euclidean embedding methods for network distance estimation in general provide reasonably accurate distance estimation for a majority of nodes,

* This work was supported in part by the new faculty research program 2006 of Kookmin University in Korea and by the NSF grant CNS-0435444.

there are fundamental limitations on their accuracy. In particular, recent studies [6, 7] have shown that the *triangle inequality violations* (TIVs) prevalent in the network distances among Internet hosts fundamentally limit the suitability of Euclidean embedding of network distances, thus the accuracy of Euclidean embedding-based distance estimation methods. Despite these limitations, Euclidean embedding methods are nonetheless appealing and useful for some applications, due to their simplicity and scalability. For example, P2P applications can easily employ the geographic forwarding such as GPSR for scalable object look up by assigning coordinates to hosts and objects based on network distances and random hash functions ([8], [9]).

In this paper, we investigate why the Euclidean embedding shows reasonable accuracy despite the prevalence of TIVs in network distances. In particular, we explore the effects of *clustering* among Internet hosts on network distance estimation – in particular, in terms of landmark selections – and how to exploit such clusters to judiciously select landmarks to obtain more accurate distance estimations. Clustering of Internet hosts are primarily due to the Internet routing hierarchy and AS (Autonomous System) topology. In other words, there are inherent clusters of hosts where distances between hosts within the same cluster are significantly smaller than those across clusters. In [7] we have showed that distances (i.e., latencies) among hosts within the same cluster tend to have more TIVs than among hosts in different clusters. In this paper based on mathematical analysis and simulation experiments, we demonstrate that reasonably good estimation of inter-cluster distances can hide the inaccuracy of small distances due to TIVs. Our findings provide us with good guidelines as to how to select landmarks to improve the accuracy of distance estimation. For instance, landmarks should be selected from each cluster and the number of landmarks should be proportional to the size of the clusters.

Before we proceed to present our work in more details, we would like to emphasize that the goal of this paper is not to try to improve the accuracy of Euclidean-embedding based distance estimation methods, which, as stated earlier, are fundamentally limited by the prevalence of TIVs. Instead, the goal is to understand the underlying factors that contribute to reasonably accurate distance estimations using the Euclidean embedding approach (while within the confines of its fundamental limitations), and to provide good guidelines for landmark selections to produce best possible results. The remainder of the paper is organized as follows. Section 2 describes the GNP and Virtual Landmarks methods. In Section 3, we show that clusters help improve the accuracy of Virtual Landmark Method. We discuss the effect of clusters on the landmark selection in Section 4. We conclude the paper in Section 5.

2 Background

In this section, we describe two representatives of Euclidean embedding based distance estimation schemes : Global Network Positioning (GNP) ([2]) and Virtual Landmarks ([3]).

GNP uses a fixed set of landmarks as the reference points. The landmarks measure the distances among themselves and assign coordinates by using simplex downhill optimization method. Basically, they assign coordinates such that the error between the actual distance and the estimated one is minimized. Then, each host measures the distances from itself to the landmarks. Based on the already assigned coordinates of the landmarks and the measured distances, each host finds out the coordinates that minimize the estimation errors by using iterative simplex downhill method.

However, the iterative simplex downhill method has very high computation time. To reduce the computation time, Virtual Landmarks method employs Principal Component Analysis (PCA). PCA is based on the singular value decomposition on the symmetric distance matrix among n nodes. The following description on singular value decomposition is mostly adopted from [3]. Let D be the $n \times n$ matrix and each entry d_{ij} is the distance from node i to node j. The singular value decomposition of the matrix D has the form

$$D = U \cdot W \cdot V^T, \tag{1}$$

where U is an $n \times n$ orthogonal matrix , V is an $n \times n$ orthogonal matrix, and W is an $n \times n$ diagonal matrix. The diagonal entries of W are called the singular values of the matrix D. The singular values of D are the nonnegative square roots of the eigenvalues of $D^T D$, and the columns of U and V are orthonormal eigenvectors of DD^T and $D^T D$. The number of non-zero singular values is the rank of the matrix D. Let $x_1, x_2, \cdots, x_k$ be the $k(< n)$ eigenvectors corresponding to the k largest eigenvalues. We stack the vectors into rows to form a transformation matrix, $M \in \mathbf{R}^{k \times n}$, i.e., $M = (x_1, x_2, \cdots, x_k)^T$. The dimension reduction is by simply multiplying M to a given high dimensional distance vector, $v \in \mathbf{R}^n$, i.e., $v' = Mv$, where $v' \in \mathbf{R}^k$.

In Virtual Landmarks, the distances among the landmarks are first measured to form the distance matrix D. Then, the transformation matrix M and the coordinates of the landmarks are computed based on the above discussion. Each host measures the distances from itself to the landmarks (let's call the distance vector, h. Such h is also called Lipschitz Embedding.) By computing Mh, the coordinates of each host are computed. One thing to note is that the number of the large singular values of D can represent the number of clusters as described in the next section.

3 Impact of Clusters on the Accuracy of Euclidean Embedding

Euclidean embedding of network distances is basically an optimization problem. It tries to assign coordinates to hosts so that the difference between the estimated distance and the real one is minimized. GNP strictly follows this idea by using simplex downhill method. Even though GNP uses the two phase coordinate computation (one for landmarks and one for hosts), which is different from the global optimization, it mimics the global optimization in such a way that the accuracy of GNP may approach the accuracy of the optimal embedding. Especially when the data set is from a Euclidean space, GNP is able to find the very accurate coordinates up to some precision errors of the machine.

Virtual Landmarks method, however, does not have any strong justification on why the Lipschitz embedding can estimate the distances with reasonable errors. All [3] shows is that the PCA can reduce the dimension of the Lipschitz embedding without much accuracy loss from the accuracy of Lipschitz embedding. This thought motivates us to investigate why PCA-based Euclidean embedding shows reasonably good estimation accuracy. We conjecture that the existence of clusters may have some impacts on the accuracy. One intuition is that when the number of clusters is small, the Lipschitz embedding can achieve good accuracy for estimating the inter cluster distances, which shows the reasonable good accuracy overall.

To justify our conjecture, we first show the estimation accuracy of Virtual Landmarks method over various synthetic and real measurement data sets. Then, we relate the accuracy with the number of clusters in the data set, which is accurately found by the number of large singular values of PCA used in Virtual Landmarks method. For the metric of the accuracy, we use the relative error, $r_{x,y}$, which is defined as follows.

$$r_{x,y} = \frac{|d_{x,y} - \hat{d}_{x,y}|}{min(d_{x,y}, \hat{d}_{x,y})}, \tag{2}$$

where $d_{x,y}$ is the actual distance between hosts x and y and $\hat{d}_{x,y}$ is the estimated one.

We first generate two types of synthetic distance matrices : Random points and Clustered points from Euclidean spaces. For the random point data sets, we randomly generate 360 points from a unit hyper cube of 2 and 8 dimensional Euclidean space. They are called "d-2" and "d-8" respectively. For the clustered points, we first select k (the number of clusters, 6 in the experiments) points as cluster centers in a unit hyper cube. Then, we generate c nodes within a small hyper cube (side length is 0.1, which is 10% of the side of the unit hyper cube. c is 60, 30, and 20 depending on the number of clusters.) centered at each cluster center. The number of dimensions is 2 and 8. We construct the distance matrix among the nodes by using the Euclidean distance between each pair of nodes. The two distance matrices are called "d-2-cl" and "d-8-cl" according to their dimensions.

Furthermore, we use two real measurement data sets : Planetlab and King. PlanetLab is derived from the distances measured among the Planetlab nodes on Sep 30th 2005 [10]. We choose the minimum of the 96 measurement data points for each measurement between node pairs. After removing the hosts that have missing distance information and the hosts that have same /24 prefixes, we obtain a 148 × 148 distance matrix among 148 nodes. King data set is the one used in [4]. After removing the hosts that have missing distance information, we finally get a distance matrix among 462 hosts.

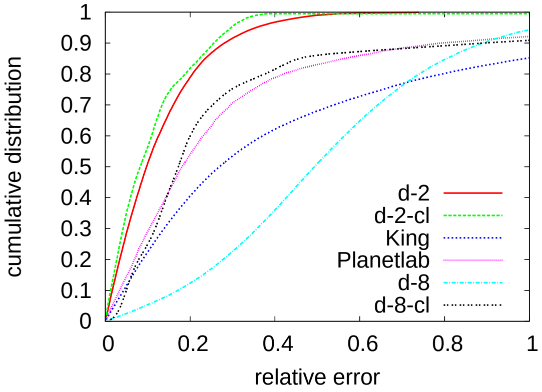


Fig. 1. Accuracy of Virtual Landmark Method : Virtual Landmarks method shows lower accuracy for high dimensional Euclidean data set. However, when the data set has clusters, the accuracy degradation is limited.

We run Virtual Landmarks method to embed the above distance matrices into the Euclidean space. In this experiment, we use 18 randomly selected landmarks. For "d-2" and "d-2-cl", we use 3 dimensions and for "d-8" and "d-8-cl", we use 9 dimensions. For King and Planetlab data sets, we use 7 dimensions, which are suggested in [2,3]. The cumulative distributions of the relative errors are shown in Fig 1. One surprising result is that the accuracy for "d-8" is very poor. However, when the data sets have clusters such as in "d-8-cl", the accuracy is reasonably good. Planetlab also shows reasonably good accuracy in that more than 70 % of the estimations show the relative error less than 0.25.

The result highly suggests that the accuracy of the Virtual Landmarks method is related with the existence of clusters. To explain this relationship, we focus on the suggestion of the authors of the Virtual Landmarks method on choosing the number of dimensions. They suggest that the number of dimensions should be the number of dominant singular values of PCA. Interestingly, in the following, we show that the number of dominant singular values is the number of clusters. This implies that the coordinates computed by the Virtual Landmarks are actually the approximate distances from each host to the clusters. [11] also states similar insights that the PCA dimension reduction automatically performs data clustering according to the K-means objective function.

To show that the number of dominant singular values is the number of clusters, we need to define the number of dominant singular values. For that purpose, we use the magnitude change $r(i)$ of the i^{th} singular value. The number of dominant singular values is defined as i such that $r(i)$ is the largest. $r(i)$ is defined as follows, where the singular values (λ_i) are sorted in descending order,

$$r(i) = \begin{cases} 1 & \text{if } i = 0 \text{ or } (\lambda_i = 0,\, \lambda_{i-1} = 0) \\ \frac{\lambda_{i-1}}{\lambda_i} & (\geq 1)\text{otherwise} \end{cases} \tag{3}$$

We first prove that the number of dominant singular values is the number of clusters for a distance matrix with extremely tight clusters, i.e., the points in each cluster are at the same position in the Euclidean space.

Theorem 1. *Let $C = \{C_1, C_2, \cdots, C_k\}$ be the k clusters of points in a d dimensional hyper cube. Each cluster C_i contains n_i points. Let $N = \sum_i n_i$. Let S be the set of N points. Let D be the $N \times N$ distance matrix among the points in S. Let λ_i be the i-th singular value of the singular value decomposition of D for $i = 0, \cdots, N-1$. If the assumption that the points in C_i are at the same position for $i = 1, \cdots, k$ holds, then*

$$r(k) = \max_{i>1} r(i),$$

for $k > 1$, where $r(i)$ is defined in (3).

Proof. Since we assume that the points in C_i are at the same position for $i = 1, \cdots, k$, the distance matrix D is $k \times k$ block matrix. The i, j block is $n_i \times n_j$ matrix. Since the points in each cluster have the same position and the distance between two nodes is the Euclidean distance, the first n_1 rows of the matrix D are the same, the second n_2 rows of the matrix D are the same, and so on. Since all the diagonal blocks are 0, there are k distinct rows in the matrix D. This means that the rank of D is k. The number of non-zero singular values of the singular value decomposition of D is actually the rank of D, i.e., k. Since $k+1$-th singular value is 0, $r(k) = \infty = \max_{i>1}(i)$. So the number of dominant singular values of D is k, the number of clusters. □

To show that the same is true for non-extreme data sets, we use three kinds of data sets including Euclidean distance matrix with clusters, Topology based synthetic distance matrices, and the real measurement data sets. For Topology based synthetic distance matrices, we first use BRITE tool from Boston University to generate synthetic 2-level topologies. Then, we move the nodes in each AS into smaller regions to make clear clusters, which look like the one in Fig. 2. We create 6 AS, 12 AS, and 18 AS topologies, and each topology has 360 nodes in total. By assigning the Euclidean distance between adjacent nodes as the weight of the link and running hierarchical routing, we compute the distance matrix among hosts. Furthermore, we use one more real measurement data set called NLANR data set. NLANR data set is collected from Active Measurement Project (AMP) ([12]) on April 7. 2004. After removing some hosts that have missing distance information, we finally get a distance matrix among 83 hosts.

We apply PCA on the distance matrices. As can be seen in Fig. 3(a), there are high peaks at the singular value number that equals the number of clusters in the Euclidean distance matrices. Similarly, Fig. 3(b) shows that the high peaks occur at the right number of ASes (i.e. clusters). This clear peak does not appear for the real Measurement data sets as can be seen in Fig. 3(c). However, there are still several reasonable peaks around 5-7, which means that Virtual Landmarks can get benefit from the existence of clusters. This is manifested in Fig. 1, where the accuracy of Planetlab is similar to that of "d-8-cl".

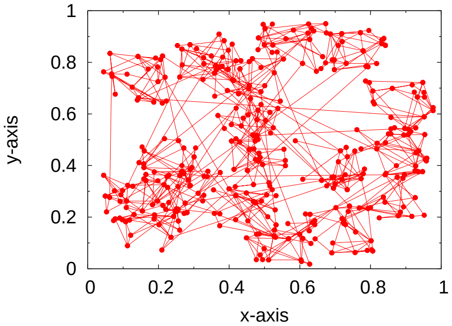


Fig. 2. Example : 18 ASes with 20 nodes in each AS. There are total 360 nodes in the topology.

To corroborate whether there exist clusters in the real measurement data set, we cluster the Planetlab data set (202 nodes including the nodes that have same /24 prefixes) by the spectral clustering algorithm[4]. Then, we compute the average intra cluster distances of the clusters and find the locations of the hosts in each cluster [5] (refer to Table 1). The average distances are computed after excluding 3 outliers (the hosts that have largest average distance to all the other hosts in the cluster) from each cluster. The second column shows the number of hosts in each cluster and the number in the parenthesis is the number of hosts after excluding 3 outliers. Most of the clusters have very small average intra cluster distances (1.328ms to 14.444ms) except the cluster 5. The cluster 5 has high intra cluster distances and the hosts of the cluster are scattered around Europe, Asia, and Australia. However the hosts in other clusters are located in relatively small regional areas. In general, the Internet has 5-7 clusters with small intra-cluster distances.

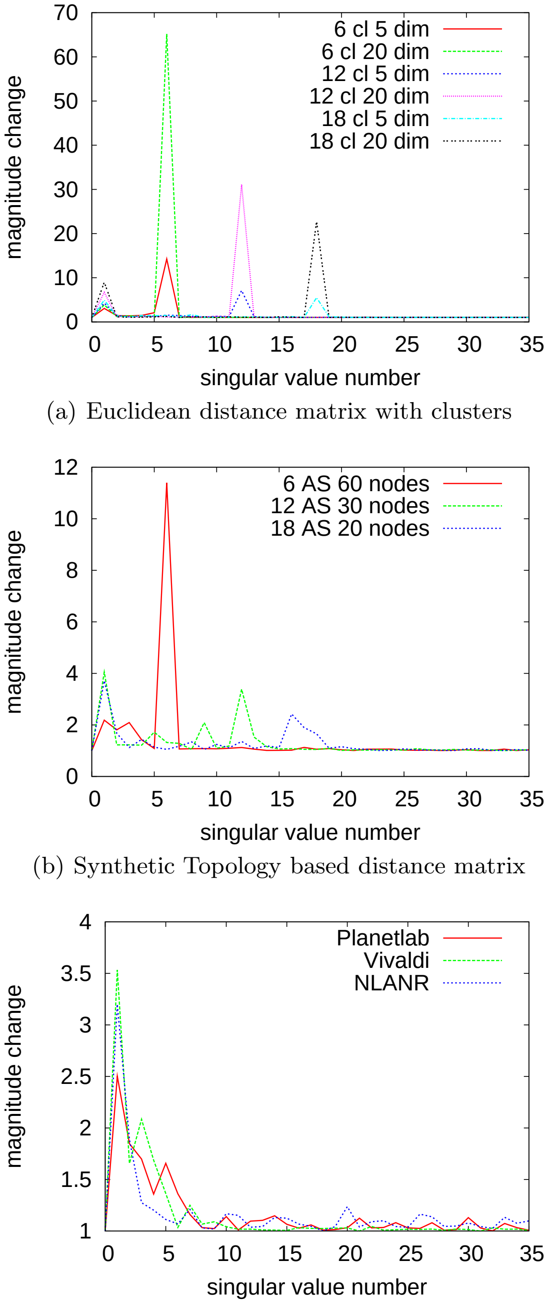
4 Effect of Clusters on Landmark Selection

In this section, we investigate the effect of clusters on the landmark selection problem. [13] shows some experiment results suggesting that one landmark from each cluster improves the accuracy. Furthermore, they show that the random landmark selection is reasonably good when the number of landmarks is around 20-30. However, the paper only shows *experimental* results rather than a rigorous analysis. Here, we provide a theorem stating that the number of landmarks selected in a cluster should be *proportional* to the number of hosts in that cluster, not just one landmark from each cluster.

Theorem 2. *Under the assumption that the hosts in each cluster are at the same position, the distance estimation that uses number of landmarks propor-*

[4] We think that any clustering method is fine for this purpose

[5] We look up the location of each IP address at "http://www.ip2location.com"



(a) Euclidean distance matrix with clusters

(b) Synthetic Topology based distance matrix

(c) Real measurement data set

Fig. 3. Magnitude changes of Principal Component Analysis

Cluster	Num of hosts	intra cluster dist (ms)	Countries	Locations
1	26(23)	14.012	USA, Canada	East cost cities,
2	31(28)	6.761	USA, Canada	East cost cities,
3	17(14)	4.322	USA	California, Washington, Arizona
4	17(14)	7.744	USA, Canada	Washington, Oregon, Calgary, Vancouver
5	56(53)	69.391	Europe, Asia, Australia	-
6	21(18)	1.328	USA	CA
7	34(31)	14.444	USA	Central USA from Michigan to Texas

Table 1. Properties of Clusters (Planet lab data).

tional to the hosts in a cluster performs better than one that uses equal number of landmarks in each cluster.

Proof. Let $\mathcal{C} = \{C_1, C_2, \cdots, C_c\}$ be the set of c active clusters. Let $\mathcal{N}$ the set of all the hosts. Let n_i be the number of hosts in cluster C_i. Let $n = \min(n_1, n_2, \cdots, n_c)$. In the proportional landmarks case, the number of landmarks in cluster C_i is $\frac{n_i}{n}$ (for simplicity, we assume all these numbers are integers). Let $\mathcal{P}$ be the set of landmarks used in distance estimation in the proportional landmarks case. Let $k = \frac{\sum_{i=1}^{c} n_i}{nc}$ be the number of landmarks per cluster used in distance estimation in the equal landmarks case. Let $\mathcal{E}$ be the entire set of landmarks used in distance estimation in the equal landmarks case. So $|\mathcal{P}| = |\mathcal{E}| = kc = \frac{\sum_{i=1}^{c} n_i}{n}$. Let $\mathcal{L}$ be the set of c landmarks, one from each cluster. So $\mathcal{L}$ is a subset of $\mathcal{P}$ and $\mathcal{E}$. We assume that the objective function of the distance estimation system is

$$\min \sum_x \sum_y |d_{x,y} - \hat{d}_{x,y}|^2, \tag{4}$$

where $x, y \in \mathcal{N}$ and $d_{x,y}$ is the actual distance between x and y, and $\hat{d}_{x,y}$ is the estimated one.

Let $K_a = \sum_x \sum_y |d_{x,y} - d^a_{x,y}|^2$ where $x, y \in \mathcal{N}$ and $D^a = (d^a_{x,y})$ is the distance matrix obtained by the distance estimation method using the landmarks from $\mathcal{P}$. That is, D^a is such that $\sum_p \sum_q |d_{p,q} - d^a_{p,q}|^2$ is minimum, where $p, q \in \mathcal{P}$. Thus, D^a minimizes

$$\sum_p \sum_q l_p l_q |d_{p,q} - d^a_{p,q}|^2, \tag{5}$$

where $p, q \in \mathcal{L}$ and l_i is the number of landmarks in the cluster to which node i belongs.

Let $K_b = \sum_x \sum_y |d_{x,y} - d^b_{x,y}|^2$ where $x, y \in \mathcal{N}$ and $D^b = (d^b_{x,y})$ is the distance matrix obtained by the distance estimation method using the landmarks from $\mathcal{E}$. That is, D^b is such that $\sum_p \sum_q |d_{p,q} - d^b_{p,q}|^2$ is minimum, where $p, q \in \mathcal{E}$.

Since all the hosts in a cluster are assumed to be at the same location, we have $K_a = \sum_p \sum_q \eta_p \eta_q |d_{p,q} - d^a_{p,q}|^2$ where $p, q \in \mathcal{L}$ and η_i is the number of hosts in the cluster to which node i belongs. Since $\eta_p = n \times l_p$, we have $K_a = n^2 \sum_p \sum_q l_p l_q |d_{p,q} - d^a_{p,q}|^2$ where $p, q \in \mathcal{L}$. Similarly, $K_b = n^2 \sum_p \sum_q l_p l_q |d_{p,q} - d^b_{p,q}|^2$ where $p, q \in \mathcal{L}$. Since D^a minimizes (5), we have $K_a \leq K_b$. □

This result applies to any embedding scheme that tries to optimize (4) including both GNP and Virtual Landmarks method. One obstacle of applying proportional landmark selection in the real situation is that $\frac{n_i}{n}$ may not be an integer. In this case, we can select *one* landmark from each cluster. To compute landmark coordinates, we can use the weighted objective function $\sum_p \sum_q n_p n_q |d_{p,q} - \hat{d}_{p,q}|^2$, where $p, q \in \mathcal{L}$. Then, to assign coordinates of a host i, we can use the weighted objective function, $\sum_q n_q |d_{i,q} - \hat{d}_{i,q}|^2$. In other words, we give a weight to the error between the host and the landmark, and the weight is proportional to the number of hosts in the cluster.

A more serious obstacle is that we do not know the clusters in advance because we do not have the distance matrix among all the hosts. However, in the following, we show that the performance of the random landmark selection with increasing number of landmarks actually converges to that of the proportional (clustering based) landmark selection. The intuition is that when the number of landmarks is large, the number of landmarks selected from each cluster is proportional to the number of hosts in the cluster.

The data set used in this experiment is the 6 ASes (clusters) with 60 nodes in each AS topology (total 360 nodes) used in the previous section. In the clustering based method, we randomly select one host from each cluster as a landmark, since we know the clusters that the hosts belong to. We select 6 such landmarks. In the sampling based method, we randomly select a subset of hosts from the set of entire hosts as landmarks. The numbers of landmarks in the sampling based method are 6, 12, 18, 24, and 30. We use 6 as the number of dimensions. After we select the landmarks, we run the Virtual Landmark method on the data set 20 times.

Fig. 4 shows the relative errors of the 20 runs at 50th, 70th, and 90th percentiles over different landmark selection method. "CL" represents the clustering based method and "SA" represents the sampling based method. The number of landmarks in each method is appended to the key. The bars show the average relative errors with min and max values of the 20 runs. As can be seen in Fig. 4, the clustering based selection shows better performance in average. Furthermore, the clustering based selection has small min-max range at each percentile, which shows the stability of the clustering based method. However, the sampling based selection has large min-max ranges for small number of landmarks. When the number of landmarks increases, the accuracy converges to that of the clustering based method. It shows that the proportional landmark selection can be achieved by using a large number of landmarks. The data sets from 12 AS 30 node topology and 18 AS 20 node topology also show similar result.

Next, we run the same experiment with the King data set. In the clustering based method, we first apply the spectral clustering algorithm to construct 10

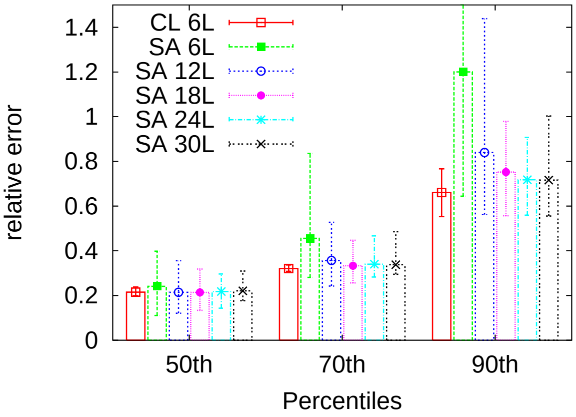


Fig. 4. Accuracy of different landmark selection methods with the Synthetic 6 AS 60 node topology.

clusters. Then, we randomly select one host from each cluster as the landmark. In the sampling based method, we randomly select a set of hosts from the set of entire hosts as landmarks. The numbers of landmarks in the sampling based method are 10, 15, 20, 25, and 30. We use 10 as the number of dimensions. We run the experiment 20 times with different sets of landmarks. Fig. 5 shows the result of the King data set. Just like the result of the synthetic data sets shown in Fig. 4, the sampling based method with 10 landmarks shows high variance on the accuracy. As the number of landmarks increases, the variance decreases, which means that the random selection approaches to the proportional selection.

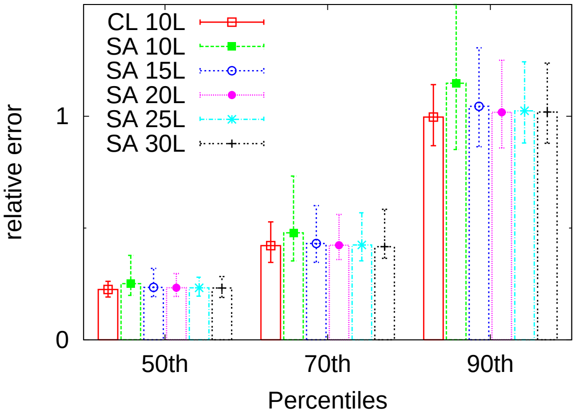


Fig. 5. Accuracy of different landmark selection methods with King data set.

5 Conclusion

In this paper, we investigated the factors that make the Euclidean embedding show reasonably good accuracy for distance estimation. We showed that the existence of clusters actually helps improve the accuracy of distance estimation in Virtual Landmarks method because of the way that the Virtual Landmarks method chooses the number of dimensions. We also showed that selecting landmarks proportional to the size of clusters increases the accuracy and in reality, the random selection of a large number of landmarks can achieve the performance of proportional landmark selection.

References

1. Francis, P., Jamin, S., Jin, C., Jin, Y., Raz, D., Shavitt, Y., Zhang, L.: Idmaps: A global Internet host distance estimation service. IEEE/ACM Transactions on Networking (2001)
2. Ng, T.E., Zhang, H.: Predicting Internet network distance with coordinates-based approaches. In: Proc. IEEE INFOCOM, New York, NY (June 2002)
3. Tang, L., Crovella, M.: Virtual landmarks for the Internet. In: Proceedings of the Internet Measurement Conference(IMC), Miami, Florida (October 2003)
4. Dabek, F., Cox, R., Kaashoek, F., Morris, R.: Vivaldi: A decentralized network coordinate system. In: Proceedings of ACM SIGCOMM 2004, Portland, OR (August 2004)
5. Madhyastha, H.V., Anderson, T., Krishnamurthy, A., Spring, N., Venkataramani, A.: A structural approach to latency prediction. In: Proceedings of the Internet Measurement Conference(IMC), Rio de Janeiro, Brazil (October 2006)
6. Zheng, H., Lua, E.K., Pias, M., Griffin, T.G.: Internet routing policies and round-trip-times. In: The 6th anuual Passive and Active Measurement Workshop, Boston, MA (March 2005)
7. Lee, S., Zhang, Z.L., Sahu, S., Saha, D.: On suitability of euclidean embedding of internet hosts. In: Proc. ACM SIGMETRICS, Saint Malo, France (June 2006)
8. Ratnasamy, S., Francis, P., Handley, M., Karp, R., Shenker, S.: A scalable content-addressable network. In: Proceedings of ACM SIGCOMM 2001, San Diego, CA (August 2001)
9. Yu, Y., Lee, S., Zhang, Z.L.: Leopard: A locality-aware peer-to-peer system with no hot spot. In: the 4th IFIP Networking Conference (Networking'05), Waterloo, Canada (May 2005)
10. Stribling, J.: Round trip time among planetlab nodes. http://www.pdos.lcs.mit.edu/ strib/pl_app/ (2005)
11. Ding, C., He, X.: K-means clustering via principal component analysis. In: Proc. of Int'l Conf. Machine Learning (ICML 2004), Banff, Alberta, Canada (July 2004)
12. NLANR: Nlanr amp data set. http://amp.nlanr.net/Status/ (2005)
13. Tang, L., Crovella, M.: Geometric exploration of the landmark selection problem. In: The 5th anuual Passive and Active Measurement Workshop, Antibes Juan-les-Pins, France (April 2004)