

# On the Internet Delay Space Dimensionality (Extended Abstract)

Bruno Abrahao     Robert Kleinberg  
Dept. of Computer Science  
Cornell University, Ithaca, NY 14850, USA

## 1 Introduction

Network latency plays a central role in the design of Internet services as their performance is sensitive to the choice of the communicating participants, among a larger set of alternatives. In light of this, coordinate-based network positioning systems have received considerable attention in the past few years [3, 6, 8]. These approaches aim at providing a compact representation of the Internet delay space (i.e., the matrix of measured round-trip latencies between Internet hosts) by modeling the network as a vector space. In this process, known as *network embedding*, each node is assigned a coordinate in a host metric space (e.g., Euclidean space) in such a way that the geometric distance between any two nodes estimates the real latency between them within a tolerable degree of error. Nevertheless, coordinate-based systems suffer from high instability, slow convergence, and disappointing accuracy, as pointed out by [4] and [5].

This research investigates a critical aspect influencing the effectiveness of coordinate-based systems in large scale networks, namely the dimensionality properties of the Internet delay space. The main component of such systems consists of an embedding algorithm for which the number of dimensions of the host metric space is a tunable parameter. Embeddings with different number of dimensions result in different degrees of accuracy, since distance matrices possess a minimum intrinsic dimensionality [3, 6, 8]. In addition, since these techniques are based on some variation of an optimization problem aimed at minimizing the prediction error, the algorithms suffer from the *curse of dimensionality* in which the algorithm's complexity increases with the number of dimensions, thereby becoming unable to deal with the overwhelming number of degrees of freedom. Finally, the convergence time, which increases with the dimensionality, affects stability, adaptability and accuracy of coordinates over time [4].

These facts motivate the following questions which are addressed in this research: to what extent is the dimensionality an *intrinsic* property of the distance matrix, defined without reference to a host metric such as Euclidean space? Is the intrinsic dimensionality of the Internet delay space approximately equal to the dimension determined using embedding techniques? If not, what explains the discrepancy? What properties of the network contribute to its overall dimensionality? In order to address these questions, one needs appropriate tools for exploring how and to what extent network properties drive the delay space dimensionality behavior. This characterization, apart from its implications to the performance of coordinate systems, is by itself a topic of practical interest as it uncovers properties and opens new questions on the nature and complexity of the network.

This research investigates all-pairs round-trip time (RTT) measurements between Internet hosts, collected via the King method and annotated with topological and geographic information, to study the effectiveness of different dimensionality metrics and to present measurements and analysis to shedding insight on the above questions. A fuller exposition of our research, together with the quantitative results omitted here for space reasons, will appear in [1]. The organization of this research announcement is as follows: Section 2 discusses the dimensionality measures we applied, and Section 3 demonstrates the applicability and effectiveness of the discussed measures in characterizing properties induced by geographic location of hosts and by the Internet AS topology. Finally, Section 4 offers our conclusions.

## 2 Dimensionality Measures

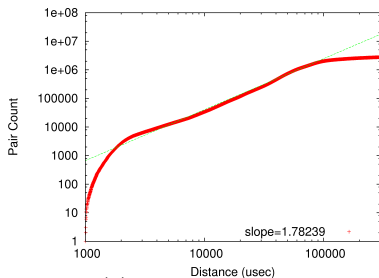
An obvious way to determine the dimensionality of a dataset is to try embedding the points in  $1, 2, 3, \dots$  dimensions using an embedding algorithm such as Vivaldi [3]. Stop at the lowest dimension  $D$  which permits an embedding with small quartiles of relative errors, and call this dimension  $D$  the *embedding dimension* of the dataset. Upon applying this process to our datasets, varying the number of dimensions from 1D to 9D, we observed that there is a drastic improvement in accuracy up to 4 dimensions and a slow improvement from 4 up to 7. Surprisingly, after 7 dimensions, the accuracy of the algorithm gets

worse, exposing the threshold beyond which the curse of dimensionality starts to affect the algorithm’s performance.

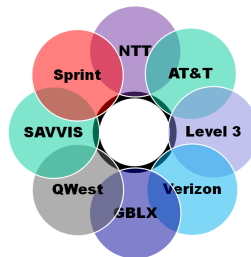
Another technique widely applied in previous work is to characterize the dimensionality of the delay space using PCA [3, 4, 8]. However, we observe that PCA does not provide a sufficiently fine-grained dimensionality measure for Internet data, and in addition, as it is a method grounded in linear algebra, when applied directly to a distance matrix it is inherently ill-suited to metric spaces which are most closely approximated by a non-linear manifold rather than a vector space.

For purposes of estimating the dimensionality of a point set there are several other ways of defining dimensionality using structural and statistical properties of the distance matrix itself, without making reference to an outside “host metric” such as Euclidean space, that capture the effect of the a richer class of geometric patterns, for instance, non-linear substructures. These measurements, which include the *correlation* and *Hausdorff* dimensions [2, 7], define the dimensionality as an intrinsic property of the distance matrix and are suitable for distance data that exhibit some form of power law behavior from either their statistical or structural properties. Point sets whose distances satisfy power laws (often with non-integer exponents) are said to behave like *fractals* [2, 7]. Accordingly, we refer to these metrics as fractal measures.

To illustrate how these measures work, the correlation dimension can be computed by plotting in logscale the number of pairs of nodes that are within a given distance from each other, for each distance in the range of interest. Figure (a) shows the pair-count plot, of the Internet delay space as represented by one of our datasets. The first striking feature of this plot is a power law that persists roughly over two orders of magnitude, i.e., from 3ms to 100ms. (Note that this range of latencies includes almost every Internet route that is not trans-oceanic.) The second observation is that the value of its dimensionality is less than 2, represented by the exponent of the pair-count curve, 1.782, which is much smaller than what the embedding dimension indicates (i.e., between 4 to 7 dimensions). The fractal measures can also help us understand the weaknesses of embedding algorithms, by showing how they corrupt the properties of the original delay space. Accordingly, upon computing the pair-count plot of the embedded network produced by Vivaldi in 7 dimensions, we observe that the resulting coordinate space loses the geometric properties (i.e., power law) of the original delay space and suffers a major dimensionality inflation.



(a) Pair-count plot



(b) Dimensionality Reducing Decomposition

In the next section we discuss why the fractal measures are effective in exploring the properties of the Internet delay space, and present our findings using them.

### 3 On the Delay Space Structure

In an attempt to understand the features of the delay space that contribute to its dimensionality behavior, we study how and quantify to what extent network properties are reflected in its delay space dimensionality. The analysis in this section also demonstrates the applicability and effectiveness of the fractal measures in capturing these properties.

We first observe that although the Internet hosts live in a sphere that can be described by two coordinates in spherical space, the value 2 found for the delay space dimensionality is not a reflection of the 2-dimensional structure of a sphere’s surface. Using the fractal methods, we observed that the pair-count plot of the geographic distances between hosts also exhibits a power law that, similarly to that of the delay space, persists for approximately two decimal orders of magnitude. As a consequence, it can

also be quantified using the correlation dimension and has value 0.897. This small value can be ascribed to the large empty spaces (i.e., oceans) and the non-uniform geolocation of nodes (i.e., dominant clusters in North America, Europe and Asia). From these results, we can also quantify that the contribution of geolocation does not fully explain the delay space dimensionality.

In order to partially explain the contribution of the other factors, we present a dimensionality shift that can be ascribed to the structural configuration of the Internet's autonomous system (AS) topology. This phenomenon can be shown by decomposing the network into intersecting Tier-1 AS networks, composed of themselves and their downstream customers and analyzing the geometry of each piece in isolation, as illustrated in Figure (b). We measured each of these subnetworks using the correlation dimension, Hausdorff dimension, embedding dimension and PCA; the quantitative results of these measurements are reported in [1] and summarized in the following paragraph.

The first observation is that the power-law behavior observed in the whole matrix was preserved over the same range of distances in the submatrices, though not necessarily with the same exponent. Second, with the exception of one of the subnetworks, all other networks exhibit correlation and Hausdorff dimensionality around 10% *smaller* than that of the whole matrix. A similar reduction is not achieved by other kinds of decomposition, for instance, by decomposing the network into pieces of smaller *cardinality*, or of smaller *diameter*. Interestingly, delay space dimensionality reduction cannot be explained by a corresponding reduction using the geographic matrix. Moreover decomposing the network into subnetworks had no effect on the dimensionality reported by the embedding and PCA methods.

As an evidence of the accuracy of the fractal measures, we used Isomap [9], a geometric dimensionality reduction technique, proposed in the Machine Learning community, which is sensitive to both linear and non-linear correlations between the dimensions. The results produced by Isomap confirm that the dimensionality of the subnetworks is indeed consistently smaller than that of the entire network in accordance with the fractal measures. This observation shows evidence of non-negligible non-linear structures in the delay space and suggests that methods based on non-linear embeddings might better reflect the structural properties of the Internet with a higher degree of accuracy than linear methods.

## 4 Conclusion

Characterizing the geometry of the delay space is critical for designing and analyzing effective coordinate-based positioning systems and sheds light on the nature of the network. This work investigates the Internet delay space dimensionality as an intrinsic property, defined without reference to an extrinsic host metric. Based on its power-law behavior, the structure of this space can be best described by fractal measures of dimension rather than by the embedding dimension or PCA. They also signify a much smaller dimensionality. Finally, using the high sensitivity of the fractal measures, we show how and quantify to what extent network properties, such as geographic location and AS topology, drive the dimensionality behavior and show evidence of non-negligible non-linear substructures in this space.

## References

- [1] B. Abrahao and R. Kleinberg. On the Internet delay space dimensionality, 2008. in preparation.
- [2] A. Belussi and C. Faloutsos. Estimating the selectivity of spatial queries using the 'correlation' fractal dimension. In *Proc. of VLDB*, 1995.
- [3] F. Dabek, R. Cox, F. Kaashoek, and R. Morris. Vivaldi: a decentralized network coordinate system. In *Proc. of ACM SIGCOMM*, 2004.
- [4] J. Ledlie, P. Gardner, and M. Seltzer. Network coordinates in the wild. In *Proc. of USENIX NSDI*, 2007.
- [5] E. K. Lua, T. Griffin, M. Pias, H. Zheng, and J. Crowcroft. On the accuracy of embeddings for Internet coordinate systems. In *Proc. of ACM/SIGCOMM Internet Measurement Conference*, 2005.
- [6] T. S. E. Ng and H. Zhang. Predicting Internet network distance with coordinates-based approaches. In *Proc. of IEEE INFOCOM*, 2002.
- [7] M. Schroeder. *Fractal, Chaos and Power Laws: Minutes from an Infinite Paradise*. W. H. Freeman and Co., NY, 1990.
- [8] L. Tang and M. Crovella. Virtual landmarks for the Internet. In *Proc. of ACM/SIGCOMM Internet Measurement Conference*, 2003.
- [9] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2000.