Understanding Geolocation Accuracy using Network Geometry

Brian Eriksson
Technicolor Research
Palo Alto, CA
brian.c.eriksson@gmail.com

Mark Crovella
Boston University
Boston, MA
crovella@bu.edu

***Abstract*—**
The ability to estimate the geographic position of a network host has a vast array of uses, and many measurement-based geolocation methods have been proposed. Unfortunately, comparing results across multiple studies is difficult. A key contributor to that difficulty is network geometry – the spatial arrangement of hosts and links. In this paper, we study the relationship between network geometry and geolocation accuracy. We define the notion of *scaling dimension* to characterize the geometry of a wide array of different networks. We show that the scaling dimension correlates with a number of aspects of geolocation accuracy. In networks with low scaling dimension, geolocation accuracy improves more rapidly with the addition of landmarks. Further, we show that the scaling dimension of operator networks varies considerably across different regions of the world. Our results point to the complexity of, and suggest standards for, the meaningful evaluation of geolocation algorithms.

I. Introduction

The ability to identify the geographic location (or *geolocation*) of Internet resources is valuable for advertisers, law enforcement, application developers, security analysts, and many others. However, resolving precise geographic information of Internet devices is limited by non-line of sight Internet routing, the lack of a standard Internet protocol providing location, and that non-mobile devices are generally not currently equipped with location identification capability (*e.g.,* GPS). Even mobile devices equipped with location services may choose not to report location information due to privacy concerns or malicious activity. In response, many approaches to geolocation have been proposed and developed. In particular, considerable attention has been paid to techniques for measurement-based geolocation, since it avoids problems of stale data and human error that plague other methods [1].

In measurement-based geolocation, one seeks to find the location of a *target* network resource represented by an IP address, using network measurements such as `traceroute` and `ping` from a set of hosts with known location, termed *landmarks.* A wide variety of measurement-based methods have been proposed [2], [3], [4], [5], [6], [7].

Evaluation of a geolocation method generally is performed using measurements taken from some set of targets, landmarks, and other hosts in the Internet. Unfortunately, comparing evaluation results across multiple studies is difficult. Typically, different studies will evaluate using hosts from different networks or regions, and using different numbers and choices of landmarks. As a result, while there are many published evaluations of geolocation methods, it can be unclear whether differences in published results are due to algorithmic performance, landmark selection, or network setting.

To illustrate this issue, Table I summarizes the test conditions used in six previous studies of measurement-based geolocation. The table shows the number of landmarks and targets, and the geographic region of the targets, across 10 experiments. Strikingly, the table shows that 9 out of 10 studies only use North American targets and 9 out of 10 studies use only a fixed number of landmarks. It is not clear what the accuracy of these prior studies would be given a different number of landmarks, or targets in another part of the world.

Algorithm Name	Landmarks	Targets	Target Locations
GeoPing [2]	14	265	US
Octant [3]	10 to 50	104	N. Amer.
Topology-Based [4]	{11,68}	{11,22,128}	N. Amer.
Constraint-Based [5]	42 95	42 95	Europe N. Amer.
Naive Bayes [6]	225	13499	N. Amer.
Street Level [7]	∼76000	{72,88}	N. Amer.

TABLE I
Measurement-Based Geolocation Studies

In this paper we take the first steps toward resolving geolocation accuracy as the number of landmarks and the geographic location of the targets change. We do so by stepping back from the geolocation problem per se, and asking what the role of *network geometry* is in measurement-based geolocation. By network geometry we mean the specific geographic location of each network node and the particular set of connections between network nodes.

Our primary tool for studying the geometry of a network is through analysis of its *scaling dimension* (closely tied to the concept of a fractal dimension [8]). Scaling dimension examines how the maximum delay to a landmark varies as a function of the number of landmarks. We show that the scaling dimension sheds considerable light on the properties of two representative geolocation algorithms: Shortest Ping [4] and Constraint-Based Geolocation [5]. We evaluate scaling dimension and geolocation accuracy (for both algorithms) for 24 real network topologies. We also evaluate scaling dimension and geolocation accuracy on two datasets consisting of measured

delays between nodes spanning multiple geographic regions.

Our contributions are two-fold. First, we show that scaling dimension informs us about how the number of landmarks affects geolocation accuracy. We demonstrate that a network with larger scaling dimension tends to show a smaller improvement in geolocation accuracy with additional landmarks, compared to a network with smaller scaling dimension. Second, we show that operator networks in different regions of the world show markedly different scaling dimension, in a way that is consistent across regions. This suggests that geolocation accuracy is likely to behave differently in separate geographic regions, but in a predictable way. Our results show that operator networks in North America, Japan, and Europe generally have high scaling dimension, while those in South America and Oceania (*i.e.,* Southeast Asia and Australia) have smaller scaling dimension. This is consistent with our related observation: North American, Japanese, and European operator networks require more landmarks to obtain a fixed percentage increase in geolocation accuracy than do South American and Oceanic networks.

Besides providing insight into the fundamental way that network geometry affects geolocation, our results also help interpret existing geolocation studies, and suggest standards for future studies.

II. Related Work

Considerable prior work has been done on the subject of measurement-based geolocation [2], [3], [4], [5], [6], [7]. We are motivated by the fact that each of these cited works has used a different geolocation data set to test their methodology, and therefore an 'apples-to-apples' comparison across different studies is not possible. Thus, rather than developing an additional geolocation technique, here we try to bridge the gap between incomparable studies using the tool of scaling dimension.

In our experiments we focus on characterizing geolocation accuracy using two different methods: Shortest Ping [4], and Constraint-Based Geolocation (CBG) [5]. We chose these two methods because they are very different, and represent two general strategies for geolocation. In Shortest Ping, the location assigned is the same as one of the landmarks, and a single landmark ultimately determines the location assignment. In contrast, in CBG, the location assigned is not that of any single landmark, but is instead the intersection of a set of feasible geographic regions with respect to *all* of the landmarks.

Not only are these two methods representative, but they are often incorporated in other geolocation methods. For example, Shortest Ping is studied in [4], but it is also used as the final step in [7]. The GeoPing method [2] assigns the target location to the 'nearest' landmark (in a synthetic delay space). The Octant algorithm relies on an augmented form of CBG to determine a feasible geographic region for a given target [3]. Additionally, CBG is used as an early step in the "Street-Level" geolocation framework in [7].

III. Datasets

We first synthesize network distance measurements through the use of 24 Internet core network topologies with geolocation information courtesy of the Internet Topology Zoo [9] and the disclosure of various tier-1 ISPs that are publicly available. Network distances are synthesized via shortest-path routing along the specified router-level network geography and connectivity.

To demonstrate the power of our analysis methodology on IP geolocation studies, we show how a collection of latency measurements from landmarks can be used to further analyze a network topology. The first dataset considered consists of 431 commercial hosts belonging to Akamai Technologies in the continental United States with known geolocation (with accuracy down to the GPS coordinates). During the weekend of January 16-17, 2010, full mesh pairwise bidirectional measurements of latency were performed between servers belonging to Akamai Technologies hosted at 431 distinct physical locations in the United States. The measurements were conducted using a single latency measurement using the MTR tool [10]. The second dataset is from a set of pairwise latency courtesy of [11] from round-trip times between 87 worldwide Planetlab nodes [12] taken in January 2008.[1]

IV. Network Geometry and Geolocation

Our focus is on the typical measurement-based approach to IP geolocation. Specifically, where there are a set of *landmarks* in the network – hosts with known location – and there are *targets*, which are to be geolocated. The user has the ability to measure delays between any target and any landmark which may be measured using ICMP ECHO, one-way probes, or other techniques, and can be processed by taking many measurements and using only the minimum. We use the term *delay* to refer to the results of any of these techniques regardless of the details.

Delay-based geolocation of a particular target t generally consists of four steps:

1) Select a set of M landmarks, $\mathcal{L}$;
2) Measure delay from target t to each landmark in $\mathcal{L}$;
3) Convert delays to distances (or distance bounds);
4) Use distances and known landmark locations to form a location estimate for target t.

Various proposals for geolocation methods address one or more of these steps. However, regardless of the algorithmic and measurement techniques used to address each step, there are fundamental limitations placed on each step by the *network geometry*. The network geometry constrains each of the four steps in standard geolocation: it restricts the locations that can be used as landmarks, it determines the geographic routing of the delay probes that can be measured, and it affects the amount and nature of geolocation information that can be extracted by algorithms.

[1]The authors would like to thank Cristian Lumezanu for making this data set publicly available.

A. Geolocation and Covering

A particularly simple approach to geolocation is "*Shortest Ping*." In Shortest Ping geolocation, delay measurements are taken between the target and landmarks, with the target geolocation assigned as the location of the landmark with the smallest delay. Given a set of M landmarks, one way to assess the accuracy of Shortest Ping is by the largest delay, denoted by r, between any node and its assigned landmark. A placement of M landmarks is equivalent to a *covering* of the network, meaning that all nodes are within r of one or more landmarks. Loosely, we can think of this covering as M "balls," each centered at a particular landmark, and having radius no larger than r (measured in units of delay in the network). For any given r, we define a *minimal covering* as a covering that uses the minimum number of landmarks necessary to cover the network with balls of radius r. We denote the number of landmarks in the minimal covering as $m(r)$. An example of a minimal network covering is shown in Figure 1-**(A)**.

A central theme in our work is to study the relationship between the covering length r, the minimum number of landmarks $m(r)$, and the average geolocation error $e(r)$. Understanding this relationship will allow us to answer the following key question: *How does geolocation accuracy scale as we increase the number of landmarks?* And although in this section these quantities will be derived from the Shortest Ping geolocation method, we later show that they shed light on more complex geolocation methods as well.

As we will show in Section V, we find that in empirical network measurements, the relationship between $m(r)$ and r can in general be approximated by a power law over a set of scales of interest $(r_{\min}, r_{\max})$:

$$m(r) \propto r^{-\beta} \qquad r_{\min} \leq r \leq r_{\max},\ \beta \geq 0 \tag{1}$$

We will show that the exact decay of $m(r)$ with r, given by the parameter β, varies for different network topologies under consideration. This decay parameter, β has a direct connection to the notion of a *fractal dimension* [8]. Our network measurements result in finite sets, which can only show scaling over a finite range of r values. For these reasons we refer to β as the *scaling* dimension of the network geometry, to distinguish it from the traditional notion of fractal dimension.[2]

B. Graph Motif Example

To illustrate the significance of scaling dimension on geolocation accuracy, consider the two simple graphs: a ring graph and a grid graph. Both graphs are embedded in two-dimensional space with the same surface area. The ring graph has scaling dimension of approximately one, since all the nodes lie on a one-dimensional line. Meanwhile, the grid graph has scaling dimension of approximately two, as all the nodes and connections are spread across two-dimensional space.

[2]Thus, while the fractal dimension is a mathematical construct pertaining to infinite objects, the scaling dimension β allows us to characterize empirical data (as discussed in [13]).

Figure 1-**(B)** shows the results for Shortest Ping geolocation on these simple graphs, each with 100 nodes. The figure shows a clear power law trend for each graph (indicated by a straight line in a log-log plot). The figure illustrates two important effects. First, for the lower-scaling-dimension ring graph, the addition of more landmarks has a significantly larger impact on geolocation error reduction than for the higher-scaling-dimension grid. Second, for small numbers of landmarks, the absolute accuracy is better for the grid graph, while for large numbers of landmarks, absolute accuracy is better for the ring graph.

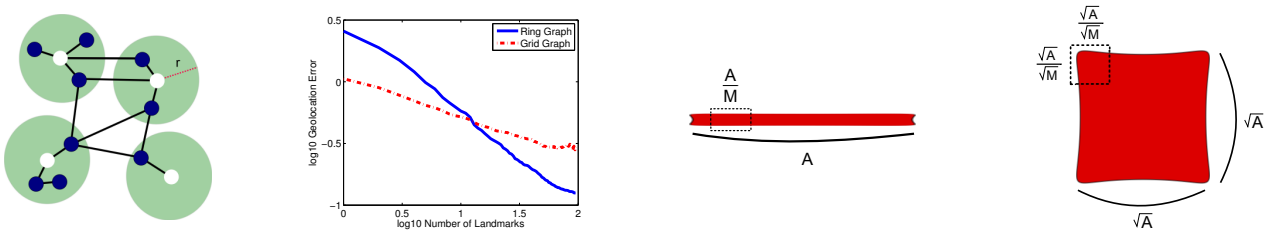
We can explain these two effects intuitively in terms of Figure 1-**(C,D)**, which depicts two "network" objects that both span a geography with area A with nodes scattered-at-random across the geography. The figure shows that using the same number of landmarks, M, in the network with scaling dimension of $\beta = 1$ nodes will typically be on the order of $O(A/M)$ away from a landmark, while in the network with scaling dimension of $\beta = 2$, nodes typically will be on the order of $O\left(\sqrt{A}/\sqrt{M}\right)$ away from a landmark. Thus, when M is small, geolocation accuracy is better for $\beta = 2$, because of the difference between A and $\sqrt{A}$ in the numerator. However, when M is large, the slower scaling of $1/\sqrt{M}$ as opposed to $1/M$ means that geolocation accuracy is better for $\beta = 1$. Generally, we would expect the performance of geolocation to scale like $O\left(M^{\frac{-1}{\beta}}\right)$, the separation distance between landmarks.

This illustrates two key ideas that are important in the rest of the paper. Consider two networks $\mathcal{N}_1$ and $\mathcal{N}_2$, such that $\mathcal{N}_1$ has larger scaling dimension than $\mathcal{N}_2$. First, $\mathcal{N}_1$ *requires the addition of more landmarks to improve geolocation accuracy by a given factor than does* $\mathcal{N}_2$. And second, $\mathcal{N}_1$ *may have better accuracy for a small number of landmarks, and poorer accuracy for a large number of landmarks, than* $\mathcal{N}_2$.

V. Experiments

Our results are obtained by applying the methods described in the last section on two kinds of data. We first analyze known topologies, annotated with geolocation information, courtesy of the Internet Topology Zoo Project [9]. Each known topology consists of a single provider's network lying within a particular geographic region (North America, South America, Japan, Europe, or Far East/Australia). Using known topologies allows us to restrict our attention to specific geographic regions, and to obtain many sample topologies from each region. We synthesize delay measurements based on the geographic length of shortest-paths. We believe that shortest-paths are a reasonable approximation for actual paths when paths are wholly contained within a single provider network, as is the case here.

Next, we demonstrate that our results hold as well when using actual measured delays over observed Internet paths. Hence, these delays are the result of true Internet routing, and the paths used extend over multiple provider networks. For this purpose, we use measurements taken from full-mesh latency



(A) **(B)** **(C)** **(D)**

Fig. 1. **(A)** - Example of a minimal network covering where $m(r) = 4$, with the landmark nodes denoted in white, **(B)** - Log-log plot of geolocation accuracy of Shortest Ping on both a ring and grid graph, each containing 100 nodes, **(C)** - Covering of an object with scaling dimension $\beta \approx 1$ and **(D)** - Covering of an object with scaling dimension $\beta \approx 2$.

probing between a set of nodes, each in a known location. We demonstrate that our results hold for two different datasets: one collected from the Akamai network, and one from Planetlab.

A. Characterizing Geometry and Geolocation

Resolving the connection between network geometry and geolocation performance is dependent on our ability to precisely characterize the scaling dimension and geolocation accuracy from observed network measurements.

While estimating the scaling dimension is clear for abstract spaces like those in Figure 1-**(C,D)**, we require a methodology for estimating the scaling dimension from observed distances in non-uniform networks. When analyzing network geometry, standard approaches such as box counting [14] present a number of limitations with respect to observed network delays. To avoid these limitations, we make use of the CLUSTER-SCALING algorithm described in [15].

To measure geolocation accuracy, we need to determine the best placement of M landmarks. For a given network, this is combinatorial in the number of possible landmark points and hence computationally infeasible. Using complete knowledge of the network geometry (*i.e.,* connectivity and geolocation), we approximate the best case geolocation performance using a *greedy* choice of M landmarks. The data generally exhibits a power law decay, where for M landmarks the geolocation accuracy decays approximately as $M^{-\gamma}$. For Shortest Ping-based Geolocation we find the mean geolocation error decay with respect to M landmarks to be proportional to, $e_{sp}(M) \propto M^{-\gamma_{sp}}$. Likewise, for Constraint-Based Geolocation and M landmarks we can similarly state, $e_{cbg}(M) \propto M^{-\gamma_{cbg}}$.

B. Scaling Dimension and Geolocation

In order to investigate the effect of scaling dimension on geolocation accuracy over a wide range of networks, we studied 24 core networks from the Topology Zoo Project [9]. We estimate the scaling dimension (β) and geolocation scaling (γ) for both geolocation algorithms for networks in varying parts of the world.

In Figure 2-**(A,B)** we show a scatterplot of each network's scaling dimension (β) versus the exponent of geolocation decay (γ) for the Topology Zoo networks. This strong relationship can be measure using the *coefficient of determination* [16], R^2, which measures the amount of variation in γ that can be explained by the variation in the expected behavior of $\frac{1}{\beta}$. For Shortest Ping-based geolocation, we find $R^2 = 0.855$ (where 85.5% of the variation in geolocation accuracy can be explained by the estimated scaling dimension), while Constraint Based Geolocation shows $R^2 = 0.787$. This demonstrates that the estimated scaling dimension can resolve a significant amount of information about the effectiveness of IP geolocation on a particular network topology.

These figures illustrates a number of important points. First, there is a consistent relationship between scaling dimension and geolocation – in general, when scaling dimension is large, geolocation accuracy tends to decline more slowly with increasing landmarks (*i.e.,* γ tends to be small). This confirms a central result, namely, that network geometry has a strong impact on geolocation accuracy, and that scaling dimension captures this impact.

Another striking characteristic of these experiments is the similarity of metrics for networks in the same continental region. This is summarized in the aggregated results in Table II. As the table shows, North American, Japanese, and European networks exhibit high estimated scaling dimension, and relatively slow decay of geolocation error with increasing numbers of landmarks. In contrast, the Oceanic and South American networks both have very low estimated scaling dimension, and much faster improvement of geolocation error as additional landmarks are employed. This suggests that geolocation algorithms are likely to show very different performance in different parts of the world.

C. Measured Delays in the Internet

We show that our results are still valid when we eliminate shortest-path assumptions. In particular, we shift to analyzing measurements taken from the Internet, in which paths cross many providers' networks, and which use actual measured delays. We note that the measurements we use here may well show triangle inequality violations. To estimate scaling dimension, we hold out a randomly chosen subset of 30 nodes as landmarks for each real world network. We then examine how the scaling dimension estimated from a full mesh probing of the 30 landmarks correlates with scaling of geolocation accuracy with respect to the remaining nodes, considering both Shortest Ping and Constraint-Based Geolocation. The results on real world networks are shown in Figure 2-**(C,D)**. The figure shows that the inverse relationship between β and γ also holds in our measurement data for both geolocation

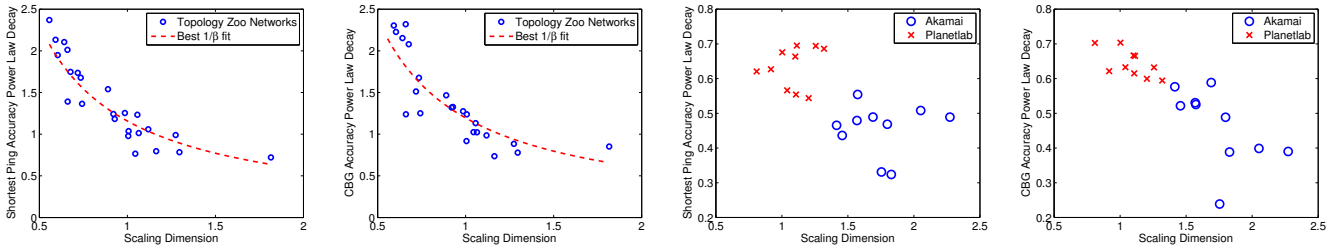


Fig. 2. Comparisons between estimated scaling dimension and greedy power law decay of the geolocation techniques. **(A)** - Shortest Ping results from Topology Zoo networks, **(B)** - CBG results from Topology Zoo networks, **(C)** - Shortest Ping results from real-world (Planetlab/Akamai) networks (each point represents results from a randomly selected set of landmarks), **(D)** - CBG results from real-world (Planetlab/Akamai) networks.

TABLE II
SCALING DIMENSION AND GEOLOCATION DECAY RATE, AGGREGATED BY GEOGRAPHIC AREA.

		Scaling Dimension Dimension (β)		Greedy Shortest Exponent (γ_{sp})		Greedy CBG Exponent (γ_{cbg})	
Geographic Area	# Networks	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Japan	2	1.104	0.083	0.780	0.021	0.880	0.204
Europe	7	1.148	0.320	0.993	0.190	1.064	0.228
North America	8	0.924	0.223	1.439	0.361	1.477	0.516
South America	3	0.681	0.053	1.620	0.420	1.547	0.524
Oceania	4	0.617	0.069	2.047	0.269	2.143	0.440

algorithms, just as it did for the topologies previously studied. The figure also shows that the two data sets have sharply different scaling dimension. We note that the Planetlab data spans multiple continental regions while the Akamai hosts are located only in the continental United States.

VI. Conclusions and Future Work

Despite considerable efforts, accurate measurement-based geolocation remains an open problem. Many proposals have been put forward for, but it is difficult to ascertain which method is best – or even, under what conditions a given method is best. In this paper we take a step toward identifying the aspects of network geometry that affect geolocation accuracy, and in so doing attempt to provide insights that help in comparing and evaluating geolocation methods. We define a property of a network which we call the network's scaling dimension. We show that the scaling dimension correlates with the degree to which adding landmarks improves geolocation, and the scaling dimension uncovers consistent differences between networks in different regions of the world.

Our analyses lead to a number of conclusions that speak to the study of measurement-based geolocation broadly. First, our results suggest that the relationship between the number of landmarks and the accuracy of geolocation varies considerably in different parts of the world. Hence, prior geolocation study results (*e.g.,* from only North America or Europe hosts) may not carry over to other parts of the world. Second, our results highlight the importance that the number of landmarks used has on geolocation accuracy.

In future work we look to examine larger topologies and how multi-region/multi-provider topologies behave with respect to scaling. We believe that further study of the scaling properties of network delays may ultimately allow a better understanding of geolocation studies, and a better basis for comparing and evaluating new geolocation methods.

References

[1] Y. Shavitt and N. Zilberman, "A Study of Geolocation Databases," *CoRR*, vol. abs/1005.5674, 2010.

[2] V. N. Padmanabhan and L. Subramanian, "An Investigation of Geographic Mapping Techniques for Internet Hosts," in *Proceedings of ACM SIGCOMM Conference*, San Diego, CA, August 2001.

[3] B. Wong, I. Stoyanov, and E. Sirer, "Octant: A Comprehensive Framework for the Geolocation of Internet Hosts," in *USENIX NSDI Conference*, April 2007.

[4] E. Katz-Bassett, J. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP Geolocation Using Delay and Topology Measurements," in *Proceedings of ACM Internet Measurements Conference*, October 2006.

[5] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-Based Geolocation of Internet Hosts," in *IEEE/ACM Transactions on Networking*, December 2006.

[6] B. Eriksson, P. Barford, J. Sommers, and R. Nowak, "A Learning-based Approach for IP Geolocation," in *Proceedings of Passive and Active Measurements Conference*, Zurich, Switzerland, April 2010.

[7] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, "Towards Street-Level Client-Independent IP Geolocation," in *Proceedings of USENIX NSDI 2011*, vol. 5, no. 5, Boston, MA, March 2011.

[8] B. Mandelbrot, *The Fractal Geometry of Nature*. New York: W. H. Freedman and Co., 1983.

[9] S. Knight, H. X. Nguyen, N. Falkner, R. Bowden, and M. Roughan, "The Internet Topology Zoo, http://www.topology-zoo.org."

[10] "The MTR Tool," http://www.bitwizard.nl/mtr.

[11] C. Lumezanu, R. Baden, D. Levin, N. Spring, and B. Bhattacharjee, "Symbiotic Relationships in Internet Routing Overlays," in *USENIX NSDI Conference*, March 2009.

[12] B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman, "PlanetLab: An Overlay Testbed for Broad-Coverage Services," *SIGCOMM CCR*, vol. 33, no. 3, pp. 3–12, 2003.

[13] B. B. Mandelbrot, "How Long is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension," in *Science*, vol. 156, 1967, pp. 636–638.

[14] K. Falconer, *Fractal Geometry*. John Wiley & Sons, Ltd., 1990.

[15] B. Eriksson and M. Crovella, "Estimation of Instrinsic Dimension via Clustering," in *Proceedings of IEEE SSP*, August 2012.

[16] L. Wasserman, *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer, May 2007.