Effectiveness of Landmark Analysis for Establishing Locality in P2P Networks

Alexander Allan and Giuseppe Di Fatta
School of Systems Engineering
The University of Reading
Whiteknights, Reading, Berkshire, RG6 6AY, UK
{siu07aja, G.DiFatta}@reading.ac.uk

***Abstract*—Locality to other nodes on a peer-to-peer overlay network can be established by means of a set of landmarks shared among the participating nodes. Each node independently collects a set of latency measures to landmark nodes, which are used as a multi-dimensional feature vector. Each peer node uses the feature vector to generate a unique scalar index which is correlated to its topological locality. A popular dimensionality reduction technique is the space filling Hilbert's curve, as it possesses good locality preserving properties. However, there exists little comparison between Hilbert's curve and other techniques for dimensionality reduction. This work carries out a quantitative analysis of their properties. Linear and non-linear techniques for scaling the landmark vectors to a single dimension are investigated. Hilbert's curve, Sammon's mapping and Principal Component Analysis have been used to generate a 1d space with locality preserving properties. This work provides empirical evidence to support the use of Hilbert's curve in the context of locality preservation when generating peer identifiers by means of landmark vector analysis. A comparative analysis is carried out with an artificial 2d network model and with a realistic network topology model with a typical power-law distribution of node connectivity in the Internet. Nearest neighbour analysis confirms Hilbert's curve to be very effective in both artificial and realistic network topologies. Nevertheless, the results in the realistic network model show that there is scope for improvements and better techniques to preserve locality information are required.**

***Index Terms*—Peer-to-Peer Networks; Landmark Clustering; Hilbert's Curve; Principal Component Analysis; Sammon's Mapping**

I. INTRODUCTION

In Peer-to-Peer (P2P) networks it can be advantageous to be aware of the geographical heterogeneity between nodes as a means of optimising load balancing, routing and search efficiency [1], [2], [3], [4]. These benefits are derived from exploiting the fact that nodes in close proximity enjoy lower communication latency.

In order to reap these rewards the P2P network needs to embed some measure of the topological distribution and the locality of its constituent nodes. Obtaining this information can be problematic as nodes do not have complete knowledge of the network from which to calculate a neighbourhood.

Landmark clustering [2] has been widely used to generate proximity information. If nodes are physically close to each other, they are also likely to experience similar latency in the communication path to selected landmark nodes.

A set of landmarks allows generating a multi-dimensional space, where each node is represented by a landmark vector, i.e. a vector of the typical communication latency to the landmark nodes. Nearby nodes in the network topology are expected to be represented by similar landmark vectors. Landmark cluster analysis allows identifying and quantifying node proximity without the detailed and global knowledge of the network topology.

Landmark spaces are typically high dimensional and techniques to map them to a 1-d space, like Distributed Hash Table (DHT) identifier spaces, have been studied [5], [6]. The general approach first calculates locality at a node via a latency vector to predefined landmark nodes throughout the network. The vector is then reduced to a 1-d index space by means of a dimensionality reduction technique employing a space filling curve known as Hilbert's curve.

It is known that Hilbert's curve possesses good locality preserving properties compared with other space filling curves [7], [8]. Among others, authors in [9] and [10] have studied Hilbert's curve and its locality preserving properties.

Yet there exists little comparison between Hilbert's curve and other general techniques for dimensionality reduction regardless of their practical applicability in the context of large-scale distributed systems.

Dimensionality reduction is a projection from a D-dimensional space onto an K-dimensional one, for $D > K$. A large number of dimensionality reduction techniques have been proposed in the litterature with different characteristics, properties and aims [11], [12]. Typically these techniques are used as pre-processing step in order to cope with the curse of dimensionality before an appropriate learning algorithm is applied to the data (typically $K \ll D$). In other applications, dimensionality reduction aims at the visualisation of multi-dimensional data ($K = 2$).

In the context of the landmark and identifier spaces in P2P systems described above, D is the number of landmark nodes and K is 1.

In this work the effectiveness of Hilbert's curve is compared with two other methods for dimensionality reduction: Principal Component Analysis (PCA) [13], [14], [15] and Sammon's mapping [16].

Principal component analysis is one of the most popular and widely used linear dimensionality reduction methods and provides the optimum projection in terms of the mean-square

 ISBN: 978-1-61208-102-1

error.

Sammon's mapping is a visualisation technique which performs a dimensionality reduction and is based on a non-linear approach.

These two techniques were chosen as they are well known examples of linear and non-linear dimensionality reduction. However, as they both require global knowledge of the data space, it would be impractical to implement them in large-scale distributed environments, like P2P systems. They rather serve as a benchmark with which to assess the quality of result produced by Hilbert's curve.

The experimental analysis is based on two network topology models. In the first topology nodes are placed on a 2d plane to generate an artificial distribution and to emphasize locality. This is used as proof of concept.

A second more rigorous simulation is based on a realistic network topology model with a typical power-law distribution of node connectivity in the Internet.

The overall goal of this paper is to provide an argument in support of the use of Hilbert's curve as a dimensionality reduction method via benchmark comparison with two other widely used techniques. An additional goal is to provide a quantitative evaluation of the locality information which is preserved after landmark vector analysis. This should lead to a better understanding of the benefits that can be expected from such a technique and of the margin for improvement.

The rest of the paper is organised as follows. Section II provides an overview of three techniques adopted to convert the landmark vectors into a locality-aware 1-d identifier. Section III describes the methodology adopted for the comparative analysis of the three methods. Sections IV and V provide the experimental results and their interpretation. Conclusive remarks are given in section VI.

II. Overview of Techniques

A. Hilbert's Curve

Hilbert's curve is a continuous fractal space-filling curve of finite granularity. Giuseppe Peano (1858-1932) discovered a densely self intersecting curve in 1890 which passes through every point in a 2-d space (and by extension in an n-dimensional hypercube) [17], [18]. This work was followed in 1891 by that of David Hilbert [19] who published his own version of the space filling curve including illustrations for construction (Figure 1). Hilbert's variant proves to have performance advantages (in terms of how well 'compact regions' of 2-d space are represented) over other space filling curves which explains its attraction as a contemporary multi-dimensional indexing method [20] [21].

The Hilbert's variant proceeds through each step replacing the U shape with an upside down Y. Each corner in the diagram represents an additional number in the sequence. As a mean of dimensionality reduction, it transforms the data from n to 1 dimension by assigning each point in space a number.

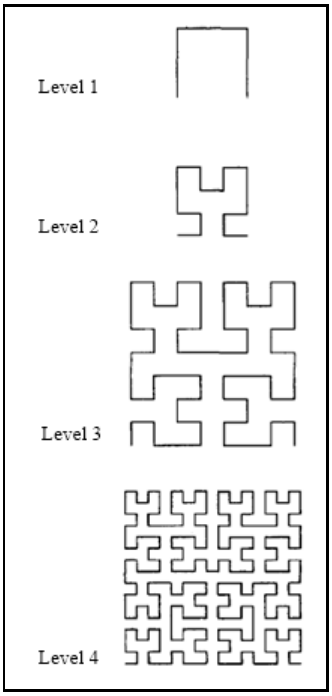

Fig. 1. The first 4 levels of Hilbert's curve in 2 dimensions

B. Principal Component Analysis

Principal Component Analysis (PCA) was introduced by Karl Pearson in 1901 [22], [15]. It employs the Karhunen-Loéve theorem which is similar to a Fourier series and transforms potentially correlated variables into a lesser number of uncorrelated variables known as principal components. PCA in essence aims to cast a projection of the higher dimensional data onto the low dimensional space from its most 'informative' angle retaining as much variability as possible.

The initial principal component chosen attempts to capture as large a range of variability within the data as is possible, with the next principal component chosen to maximize the remaining variability and so forth until all principal components are identified.

C. Sammon's Mapping

Developed by J. W. Sammon Jr in 1969 [16], Sammon's mapping is a non-linear form of dimensionality reduction based on gradient search which attempts to keep as much of the structure of the original measurement of dimensions as possible. Each iteration attempts to minimize an error function known as Sammon's stress, while matching the pairwise distances in the high-dimensional space to those in the lower-dimensional one.

PCA and Sammon's mapping have been shown to be among the best methods for dimensionality reduction in terms of preserving cluster validity [23] for data visualisation ($K = 2$).

III. Comparative Analysis

This work provides a comparative analysis of the three techniques described above for reducing dimensionality to

 ISBN: 978-1-61208-102-1

generate a locality aware index ($K = 1$) for the nodes of P2P overlay networks. A randomly generated index for control and an ideal index obtained from global network knowledge are included in the comparison for reference.

A. Simulation 1: landmark analysis on a 2d plane

In the first experiment an artificial network topology is considered as proof of concept. 1000 points were arranged on a 2d Euclidean plane following a rectangular perimeter to represent nodes on a network (Figure 2). Six landmarks are randomly selected, based on preliminary work suggesting that this number of landmarks produces the best accuracy within a range restricted by computational resources available. The Euclidean distance from a node to a landmark is used to simulate network latency and provides a 6d vector for every node. This vector was reduced to the 1d node index using all three methods of dimensionality reduction. The Hilbert number $H(n)$ was computed with a recursion free version [24][25] of the Hilbert's curve algorithm. Node indices $P(n)$ and $S(n)$ were calculated by means of the implementations of the algorithms, respectively, PCA and Sammon's mapping, available in the data mining development environment *KNIME* [26]. As a control, a random index R(n) was created in which there was no relation between index values and location.

For each network node n, the 10 nearest neighbours ($nn_1, ..., nn_{10}$) were found by searching the 10 closest indices on the 1d space defined by, respectively, node indices $H()$, $S()$, $R()$ and $P()$. The Euclidean distances were determined from node n to its nearest neighbours ($nn_1, ..., nn_{10}$) on the original 2d plane. The sum (N) of these distances is computed for each node to create distance arrays $N_H[]$, $N_R[]$, $N_S[]$ and $N_P[]$. The value of N was also found for the nearest neighbours of every point on the original 2d plane to produce an ideal neighbour value array $N_I[]$. These 5 arrays were then compared with each other to assess the degrees of locality preservation.

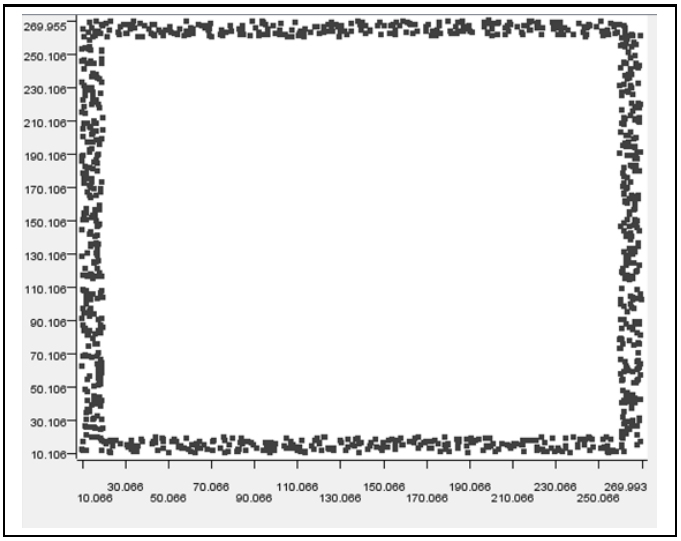

Fig. 2. Layout of the artificial network topology with 1000 nodes (2d plane)

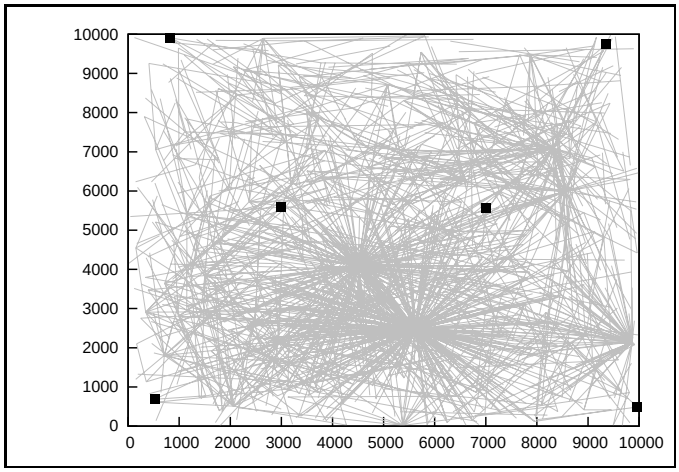

Fig. 3. A 2d representation of the realistic network topology (Inet) with 3037 nodes. (Landmarks are marked)

B. Simulation 2: landmark analysis in weighted graph

The second experiment applies the same methodology to a more realistic network topology generated with Inet [27], an Autonomous System level Internet topology generator. Inet generates random networks with characteristics similar to those of the Internet in terms of the power-law distribution of node connectivity. It can approximate Internet-like network topologies to a high degree of accuracy [28] with respect to the degree distribution and the minimum vertex cover size. Inet was used to create a 3037 node graph with 4788 edges whose weights corresponds to latency values. 6 nodes of this graph were chosen as landmarks (Figure 3).

The shortest path between pairs of nodes was computed with the Dijkstra algorithm to determine their communication latency. For each network node a 6d vector was generated with the communication latency to the landmarks. From this vector the indices H, P and S were computed and the random control R was generated. The 5 nearest neighbours for every node in each 1d space defined by the indices H, P, S and R were found and the sum of distance to these neighbours via Dijkstra shortest path was calculated. Similarly to the previous case, the latency arrays $N_H[]$, $N_R[]$, $N_S[]$ and $N_P[]$ were computed. The actual 5 nearest neighbours of every node in the network were found using a Dijkstra algorithm with expanding search radius. The total sum of the path length from each node n to its 5 nearest neighbours ($nn_1, ..., nn_5$) was computed to produce an ideal sum of latencies array $N_I[]$.

IV. Experimental Results

The landmark-vector analysis on the artificial 2d plane topology showed that Hilbert's curve is the most effective method of preserving locality information among the three dimensionality reduction techniques. Table I and Figure 4 show the result for this case. Points indicated as adjacent by the Hilbert index were on average over twice (2.13) as far as the best possible as given by the ideal case but over 19 times closer than if they had been chosen at random. Sammon's mapping came second scoring four times worse than Hilbert's

TABLE I
AVERAGE DISTANCE TO THE 10 NEAREST NEIGHBOURS FOR THE ARTIFICIAL NETWORK TOPOLOGY (2D PLANE)

Method (6 landmarks)	Mean distance to 10 nearest neighbours
Ideal	44.52
Hilbert	94.72
Sammon	404.98
PCA	883.35
Random	1810.78

TABLE II
AVERAGE LATENCY TO THE 5 NEAREST NEIGHBOURS FOR THE REALISTIC NETWORK TOPOLOGY (INET)

Method (6 landmarks)	Mean latency to 5 nearest neighbours
Ideal	269.07
Hilbert	600.54
Sammon	714.34
PCA	744.80
Random	751.18

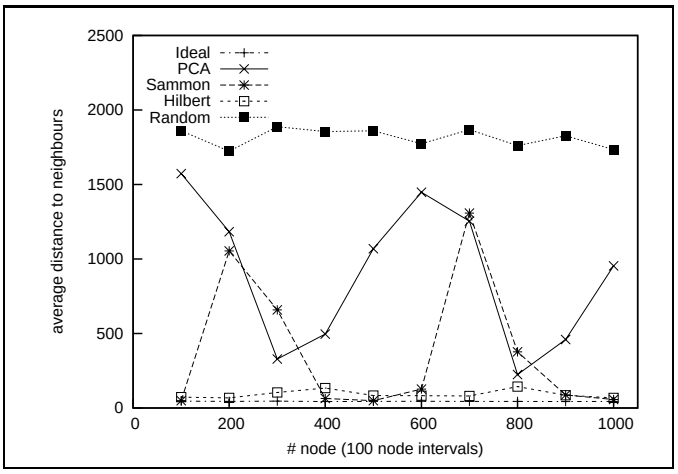

Fig. 4. Neighbour distance (moving average) for the artificial network topology model (2d plane)

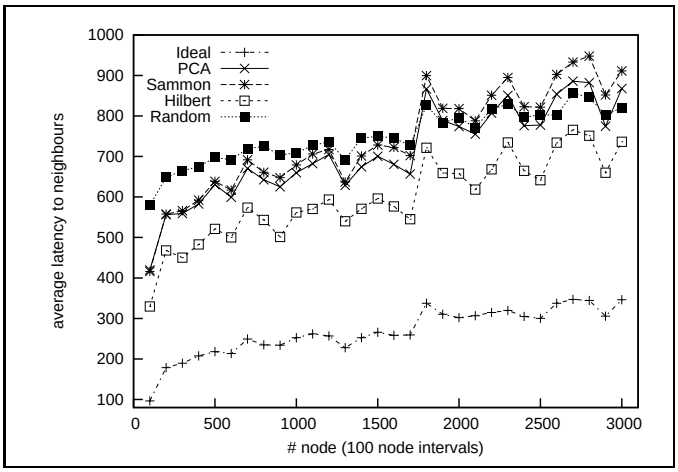

Fig. 5. Neighbour latency (moving average) for the realistic network topology (Inet). Five nearest neighbours are considered for each node

curve, but four times better than random points. PCA scored nine times worse than Hilbert's curve, but still over twice as well as random.

Table II and Figure 5 show a comparison of the different techniques for the more realistic hierarchical topology. The Hilbert's curve confirmed to be the most effective technique with an average latency of just over twice the ideal possible value (2.23 times higher) but with much less improvement over random (1.25 times lower). The PCA approach came second with just 1.05 times smaller latency to neighbours than to neighbours chosen randomly. Sammon's mapping performed the worst showing barely any improvement over random (only 1.008 times smaller on average).

In the artificial 2d plane topology an ideal nearest neighbour distance is on average 40.67 times less than to an index based on randomly chosen neighbours. In this case Hilbert's curve is very close to the ideal index. In the Internet-like topology the ratio between ideal and random indices dropped to 2.79. However, in this more realistic case the performance of Hilbert's curve is far from ideal.

V. DISCUSSION

The results on the 2d plane network model showed good performance for the landmark analysis technique as might be expected when Euclidean distance can be taken to landmarks (as this is the principle of GPS navigation systems). The Hilbert's curve produced results that were much closer to the ideal values than to the random control values. The other two techniques also performed much better than random on this test, but were less effective than Hilbert's curve.

In order to generate a homogeneous Hilbert index across independent nodes, all that is required at each node is the maximum possible size of all input dimensions and a curve width. These can be predetermined constants. For Sammon's mapping, the entire data set would need to be recalculated every time a new node joined the network as the maps produced lack generalisation. Similarly, PCA relies on knowing the entire data set to establish the covariance matrix which is central to the algorithm. For this reason, Sammon's mapping and PCA are wholly unsuitable for distributed geographical indexing purposes. In this test they really serve as a benchmark with which to assess the effectiveness of the Hilbert's curve. These methods are better suited to feature preservation than locality preservation.

In general, the results were not unexpected considering Hilbert indexing is known as one of the best methods of preserving locality, but the degree to which it outperforms other dimensional scaling methods was surprising.

The results from the more realistic topology model indicate that landmark vector analysis via Hilbert's curve can give nodes an awareness of locality. The results also show that Sammon's mapping and PCA are not likely to produce useful results, performing only marginally better than points chosen at random. Considering the chart in Figure 5, the PCA and Sammon's mapping may actually perform worse than random for some nodes (1800 to 3000).

What these results are not able to show for certain is

 ISBN: 978-1-61208-102-1

whether the relatively poor performance on the Internet-like network topology when compared to the 2d plane topology is due to the loss of landmark quality when a path is found through a network or whether this degradation is due to the dimensionality reduction. However given that the performance on the 2d plane, it seems likely that a real network represents a more complicated space through which to derive locality so performance is not likely to approach that of a more theoretical realm.

Whether the 25% improvement over random achieved by the Hilbert reduction of landmark vector analysis represents enough locality information to be useful is subject to the individual requirements of the desired application. These results give at least some indication as to the quality of accuracy likely to be achieved by using this, or similar techniques across a real network.

VI. Conclusion

This work has presented a comparative analysis of multi-dimensional scaling techniques for establishing node locality in P2P networks. A set of shared landmarks can be adopted by each node to incorporate locality information in its peer identifier. The experimental analysis on both an artificial topology model and on a hierarchical topology based on a realistic Internet model, has shown the effectiveness of the Hilbert's curve with respect to Principal Component Analysis and Sammon's mapping. Nevertheless, the analysis has also identified scope for improvements; locality preserving techniques based on Hilbert's curve and landmark clustering are far from ideal. The space filling curve used in the H-indexing scheme proposed by Niedermeier et al. [10] purports to outperform Hilbert's curves in terms of locality preservation. These curves shall be incorporated and evaluated in future implementation. Current research efforts are focusing on the optimisation of the simulation code in order to extend the analysis to larger networks, varying number of landmark nodes and neighbours. These factors are currently restricted by computational resources. To provide more realistic results still, an implementation on PlanetLab P2P overlay test bed [29] is planned. A further interesting research direction is the adoption of techniques to cope with missing latency measurements to some landmarks and to introduce robustness to variability of the set of landmarks over the network nodes.

References

[1] M. Castro, P. Druschel, Y. C. Hu, and A. Rowstron, "Exploiting network proximity in peer-to-peer overlay networks," Microsoft Research, Cambridge, England, Tech. Rep. MSR-TR-2002-82, May 2002.

[2] S. Ratnasamy, M. Handley, R. Karp, and S. Shenker, "Topologically-aware overlay construction and server selection," in *Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies Proceedings*, New York, USA, Jun. 2002, pp. 1190–1199.

[3] Y. Zhu and Y. Hu, "Towards efficient load balancing in structured p2p systems," in *18th International Parallel and Distributed Processing Symposium (IPDPS'04)*, Santa Fe, USA, Apr. 2004, p. 20a.

[4] H. Shen and C. Xu, "Hash-based proximity clustering for load balancing in heterogeneous dht networks," *Journal of Parallel and Distributed Computing*, vol. 65, no. 5, pp. 686–702, May 2005.

[5] Z. Xu, C. Tang, and Z. Zhang, "Building topology-aware overlays using global soft-state," in *Proceedings of the 23rd International Conference on Distributed Computing Systems*, Brown University, USA, May 2003, pp. 500–508.

[6] Z. Xu, M. Mahalingam, and M. Karlsson, "Turning heterogenity into an advantage in overlay routing," in *Proceedings of IEEE INFOCOM*, San Francisco, USA, Mar. 2003, pp. 1499 – 1509.

[7] C. Gotsman and M. Lindenbaum, "On the metric properties of discrete space-filling curves," *IEEE Transactions on Image Processing*, vol. 5, no. 1, pp. 794–797, Jan. 1996.

[8] B. Moon, H. Jagadish, C. Faloutsos, and J. Saltz, "Analysis of the clustering properties of the hilbert space-filling curve," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 1, pp. 124–141, Jan. 2001.

[9] J. Alber and R. Niedermeier, "On multi-dimensional hilbert indexings," in *Computing and Combinatorics: 4th Annual International Conference, Proceedings*, Taipei, Taiwan, Aug. 1998, pp. 329–338.

[10] R. Niedermeier, K. Reinhardt, and P. Sanders, "Towards optimal locality in mesh-indexings," in *Proceedings of the 11th International Symposium on Fundamentals of Computation Theory*, Krakow, Poland, Sep. 1997, pp. 364 – 375.

[11] M. A. Carreira-Perpinan, "A review of dimension reduction techniques," Department of Computer Science, University of Sheffield, U.K., Tech. Rep. CS-96-09, 1997.

[12] I. K. Fodor, "A survey of dimension reduction techniques," Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA, Tech. Rep. UCRL-ID-148494, 2002.

[13] P. Diaconis and M. Shahshahani, "On nonlinear functions of linear combinations," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, pp. 175–191, Mar. 1984.

[14] G. Eslava and F. H. C. Marriott, "Some criteria for projection pursuit," *Statistics and Computing*, vol. 4, no. 1, pp. 13–20, Mar. 1994.

[15] I. Jolliffe, *Principal Component Analysis*, 2nd ed., ser. Springer Series in Statistics. New York, USA: Springer, 2002, no. XXIX.

[16] J. W. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, vol. C-18, no. 5, pp. 401–409, May 1969.

[17] G. Peano, "Sur une courbe, qui remplit toute une aire plane," *Mathematische Annalen*, vol. 36, no. 1, pp. 157–460, 1890.

[18] A. R. Butz, "Space filling curves and mathematical programming," *Information and Control*, vol. 12, pp. 314–330, 1968.

[19] D. Hilbert, "Ueber die stetige abbildung einer line auf ein flchenstck," *Mathematische Annalen*, vol. 38, pp. 459–460, 1891.

[20] C. Faloutsos and Y. Rong, "Spatial access methods using fractals: Algorithms and performance evaluation," University of Maryland, Maryland, USA, Tech. Rep. UMIACS-TR-89-31, Mar. 1989.

[21] H. Jagadish, "Linear clustering of objects with multiple attributes," *ACM SIGMOD Record*, vol. 19, no. 2, pp. 332–342, Jun. 1990.

[22] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 6, pp. 559–572, Jun. 1901.

[23] D. Marghescu, "Evaluating the effectiveness of projection techniques in visual data mining," in *Proceedings of the Sixth IASTED International Conference on Visualization, Imaging, And Image Processing*, Palma de Mallorca, Spain, Aug. 2006, pp. 186–193.

[24] A. Butz, "Alternative algorithm for hilbert's space-filling curve," *IEEE Transactions on Computers*, vol. 20, no. 4, pp. 424–426, Apr. 1971.

[25] D. Moore. Fast hilbert curve generation and sorting and range queries. Rice University. Texas, USA. [Online]. Available: http://www.tiac.net/∼sw/2008/10/Hilbert/moore/hilbert.c (last accessed: June 2010)

[26] M. Berthold, N. Cebron, F. Dill, G. Di Fatta, T. Gabriel, F. Georg, T. Meinl, P. Ohl, C. Sieb, and B. Wiswedel, "KNIME: The Konstanz Information Miner," in *Proceedings of the Workshop on Multi-Agent Systems and Simulation MAS&S, 4th Annual Industrial Simulation Conference (ISC)*, Palermo, Italy, Jun. 2006, pp. 58–61.

[27] J. Winick and S. Jamin, "Inet-3.0: Internet topology generator," University of Michigan, Michigan, USA, Tech. Rep. CSE-TR-456-02, Jun. 2002.

[28] G. Di Fatta, G. L. Presti, and G. L. Re, "Computer network topologies: Models and generation tools," Consiglio Nazionale delle Ricerche, Palermo, Italy, Tech. Rep. CERE-CNR, 5/2001, Jul. 2001.

[29] B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman, "Planetlab: an overlay testbed for broad-coverage services," *IEEE Transactions on Computers*, vol. 33, no. 3, pp. 3–12, Jul. 2003.

 ISBN: 978-1-61208-102-1