Application of Locally Linear Embedding to the Spaces of Social Interactions and Signed Distance Functions

Kush R. Varshney

Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.881: Representation and Modeling for Image Analysis

May 5, 2005

1. Introduction

When dealing with high-dimensional data, dimensionality reduction is an important operation to allow the discovery of simple relationships between data points and simple modes in which the data varies. The input to a dimensionality reduction procedure is data in high dimension and the output is a mapping of that data to a low-dimensional manifold. Dimensionality reduction should preserve neighborhoods, *i.e.* points close to each other in the high-dimensional space should remain close in the computed low-dimensional space. Additionally, it is desirable for a dimensionality reduction method to allow nonlinear low-dimensional manifolds. Nonlinear dimensionality reduction by locally linear embedding (LLE) proposed by Roweis and Saul [1] is such a procedure that is intuitive, simple to implement, and does not involve local minima in its optimization. If the output low-dimensional manifold has two dimensions, intuitively this algorithm connects data points in high-dimensional space to neighboring data points with springs and then presses the spring structure between two glass plates [2]. Another informal description of the algorithm is that it cuts out locally linear swatches of the data in high dimension with scissors and places the swatches down in low dimension in such a way that preserves angles between data points close to each other [3], *e.g.* deconstructing a soccer ball into its constituent pentagonal and hexagonal patches and laying them out on a table after possibly deforming them.

In [1], LLE is applied to documents and to face images, but the technique is restricted neither to a particular application domain nor to a particular domain of data representation within an application. Of key significance is the wide ranging applicability of LLE, because analysis of multivariate data is a universal problem in the sciences. In the work presented here, two application areas will be explored through LLE. After reviewing the algorithm, first document analysis with a social network interpretation will be considered using data from an epic poem. Second, the space of signed distance function (SDF) representations of images, often used in segmentation, will be looked at through topographic and bathymetric data. Some conclusions will also be given.

2. Locally Linear Embedding Algorithm

In this section, the details of the simple LLE algorithm will be given, following from [3]. We start with N data points represented as D-dimensional, real-valued vectors $\mathbf{x}_i$. The first step of the algorithm is to identify the K nearest neighbors using Euclidean distance in the D-dimensional space for each data point $\mathbf{x}_i$. The second step is to determine coefficients $w_{i,j}$ that best reconstruct each $\mathbf{x}_i$ using a linear combination of the K neighbors $\mathbf{x}_j$, where best is in a squared distance sense. This is the 'locally linear' part of the algorithm. Specifically, the following cost function is minimized with respect to all $w_{i,j}$:

$$J_W = \sum_{i=1}^{N} \left| \mathbf{x}_i - \sum_{j=1}^{K} w_{i,j} \mathbf{x}_j \right|^2 \tag{1}$$

under the constraint that for each i, the sum of the K weights $w_{i,j}$ be unity. The problem is of the constrained least squares variety and has a simple closed form solution.

Once these weights are obtained, the final step of the algorithm can be performed. The goal of dimensionality reduction is to map the D-dimensional $\mathbf{x}_i$ to d-dimensional $\mathbf{y}_i$, where $d < D$. Thus with the calculated weights fixed, a cost function of the same form as (1) is minimized over the low-dimensional coordinates for $\mathbf{y}_i$:

$$J_Y = \sum_{i=1}^{N} \left| \mathbf{y}_i - \sum_{j=1}^{K} w_{i,j} \mathbf{y}_j \right|^2 \tag{2}$$

with the following constraints. The value of the above cost function (2) remains unchanged for any translation of the entire set of vectors $\mathbf{y}_i$, so one constraint is that the coordinates be centered at the origin. The cost function can be minimized by setting all of the $\mathbf{y}_i$ to equal the zero vector. To avoid this degenerate solution, another constraint is imposed: the sample covariance of the vectors $\mathbf{y}_i$ is set to unity. The optimal low-dimensional mapping can then be calculated in closed form by solving an eigenvalue problem for eigenvectors corresponding to the smallest $d + 1$ eigenvalues. The LLE algorithm is no more complicated than these three steps.

3. Mahābhārata

Documents of text are not simply a 'bag of words,' but have much structure at various resolutions, including at the sentence level, the paragraph level, and the section level. When the document is a narrative, with different sections corresponding to different scenes or events, the co-presence of two characters indicates social interaction between them. Thus, when a document is a narrative with numerous characters, analysis of the text may also lead to a characterization of the social network among the characters, by regarding characters to be close if they appear nearby in the text. Performing this sort of analysis, however, makes sense only if there are hundreds or thousands of characters. In actual networks of social interaction, the requirement of having a large number of people is met without difficulty, but there are few texts with hundreds, let alone thousands, of characters. The Mahābhārata, an epic poem popularly touted as the longest poem in the world with about 110,000 verses, does have thousands of characters and will be used as a source of data. It will be treated as a surrogate for actual social interaction data.

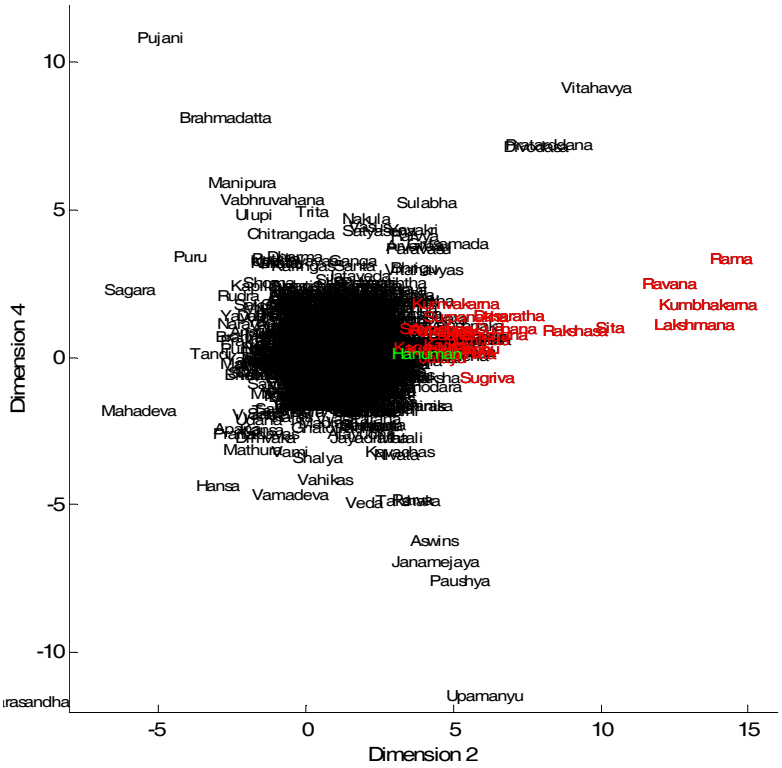


Figure 1: Rāmāyaṇa characters in the low-dimensional manifold.

The Mahābhārata is in the Sanskrit language, but an English translation [4] is used in this work for two primary reasons. First, it is significantly easier to automatically identify proper nouns in English than Sanskrit because English has capital and lowercase letters. Second, nouns in English are not declined based on their function within a sentence, remaining invariant to usage. In this work, names of individual people, of groups of people, and of places will be considered.

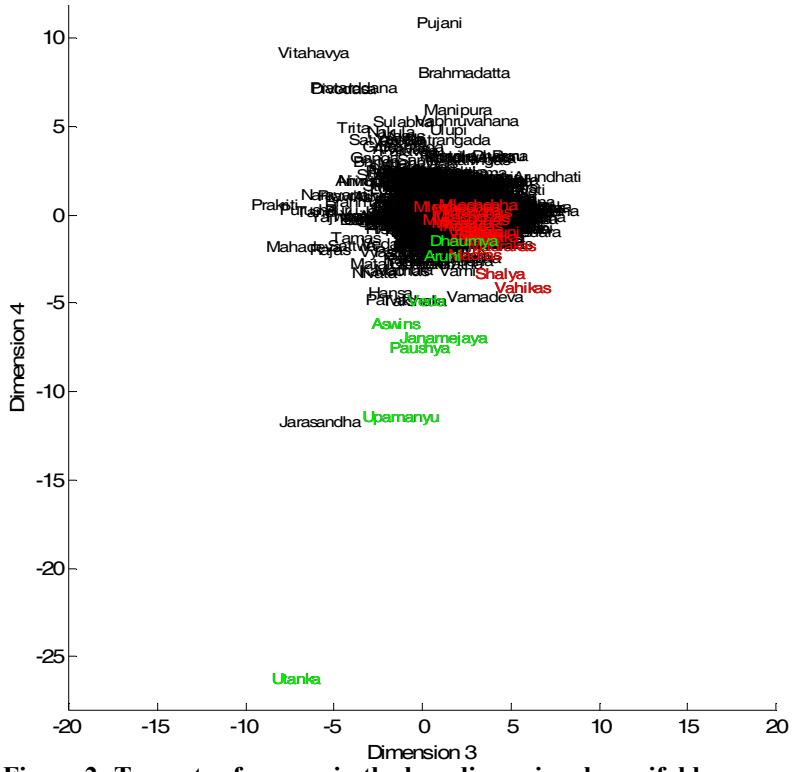


Figure 2: Two sets of groups in the low-dimensional manifold.

The epic is divided into eighteen books, each subdivided into varying numbers of sections. Overall, there are 2,110 sections, which will be treated as the events of social interaction. In the English translation, there are 10,854 distinct capitalized words, of which 7,296 were identified as proper nouns. Of the 7,296 proper nouns, 3,980 appear just once and 3,316 appear more than once. These 3,316 names are the data points that will be used in a 2,110 dimensional space. For each dimension (section), the data is a count of how many times the name appears in that section. This type of data is very similar to the text example with encyclopedia articles in [1]. Some results are presented below with the number of neighbors $K = 24$.

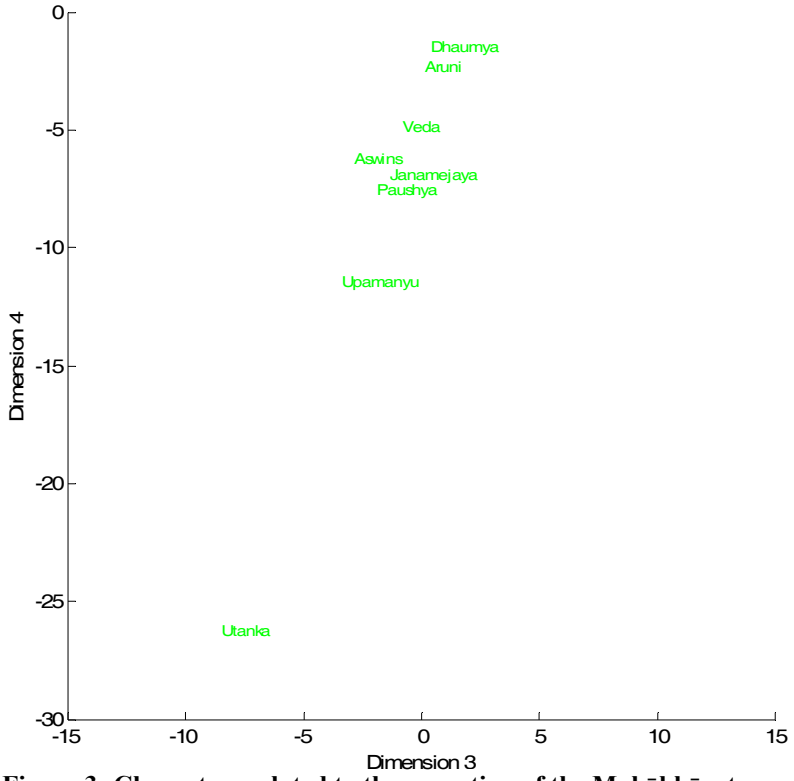


Figure 3: Characters related to the narration of the Mahābhārata.

In Fig. 1, the data points, the names, are plotted as they have been mapped to dimensions two and four of the computed

low-dimensional space. The figure illustrates that LLE is neighborhood preserving and maintains global properties as follows. The names highlighted in red, Rama, Ravana, Kumbhakarna, Lakshmana, Sita, Sugriva, Surpanakha, etc. are all characters of another epic poem, the Rāmāyaṇa, that precedes the Mahābhārata chronologically and is recounted at various points by various characters, but is not part of the actual Mahābhārata story. Not surprisingly, all of these names are kept close together in the low-dimensional space. A main character in the Rāmāyaṇa, Hanuman, is immortal and is encountered by the protagonists of the Mahābhārata in one instance. Correspondingly, the Hanuman data point, highlighted in green, is closer to the central mass of points, which are connected with the central narrative.

In Fig. 2, dimensions two and three are plotted, with two groups of names highlighted. One set is shown in detail in Fig. 3 and the other in Fig. 4. The names in Fig. 3, in green, are not characters in the main Mahābhārata story either, but lead to the story being narrated. Dhaumya is a teacher with students Aruni, Veda, and Upamanyu. Dhaumya tests Upamanyu with the help of the Aswins. Utanka is a student of Veda who is tested with the help of Paushya. Utanka goes on to prompt the telling of the

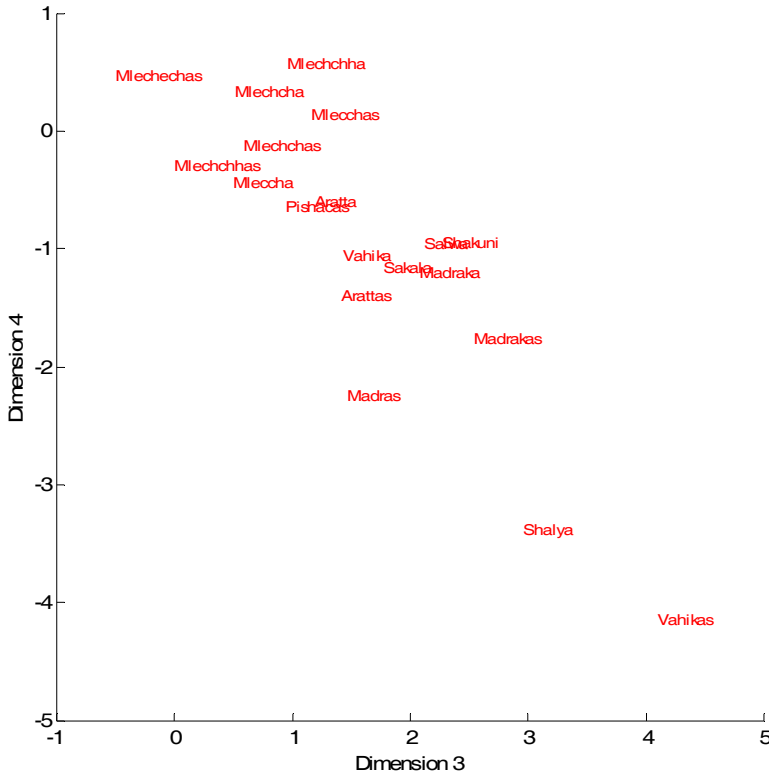


Figure 4: Places, groups, and people from the west.

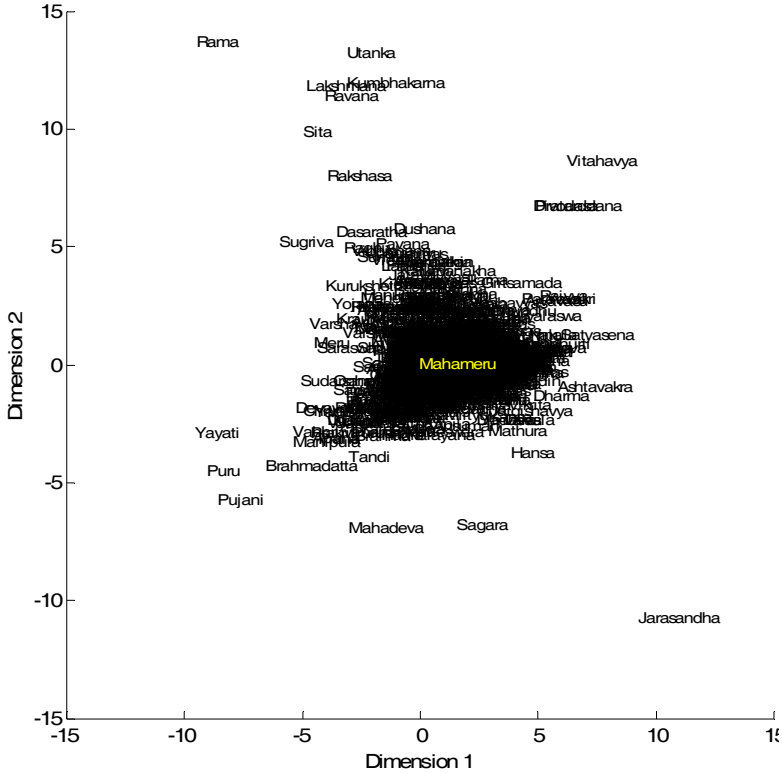


Figure 5: First two dimensions of the low-dimensional manifold.

Mahābhārata to Janamejaya. These names are also separated from the main conglomeration and fall along one axis.

The English translation is inconsistent in the transliteration of Mleccha, a group of people from the west in Persia and Greece, but nevertheless, all of the different instances fall in the low-dimensional manifold near each other and near others originating in the west in Fig. 4 in red. Salwa and Shakuni are both kings from the west. The Pishacas, Arattas, Vahikas, and Madrakas are all people from the west, from places such as Afghanistan. The Madrakas are from the kingdom of Madra, with capital Sakala and king Shalya. This grouping occurs despite the fact that Salwa, Shakuni, and Shalya have extremely different roles in the story and are not co-present much of the time. Thus LLE is able to discover modes of variation in the data that are not readily apparent otherwise, in this case, direction of origin. All of these names are important to the story and thus are part of the central mass of points.

An interesting feature of the computed low-dimensional space is the data point mapped closest to the origin in the first two dimensions shown in Fig. 5, Mahameru. It is a mountain that is purported to be the cosmic axis or the center of the physical and metaphysical universe. It is mapped close to the origin in the other dimensions as well.

Thus, this dataset of the Mahābhārata text, as a surrogate for social interaction data, illustrates that the use of LLE preserves both local and global structure. Even when the form of the structure is not known in advance, by viewing the low-dimensional space, it is possible to find properties within the dataset that group data points and separate data points.

4. Earth Relief

The signed distance function is a powerful way to implicitly represent a curve, often used in segmentation algorithms [5]. (A curve partitions a plane into two regions, with one designated the inside and the other the outside.) One intuitive way to describe signed distance functions is through analogy with Earth elevation, specifically the Hawaiian Islands [6]. Elevation data

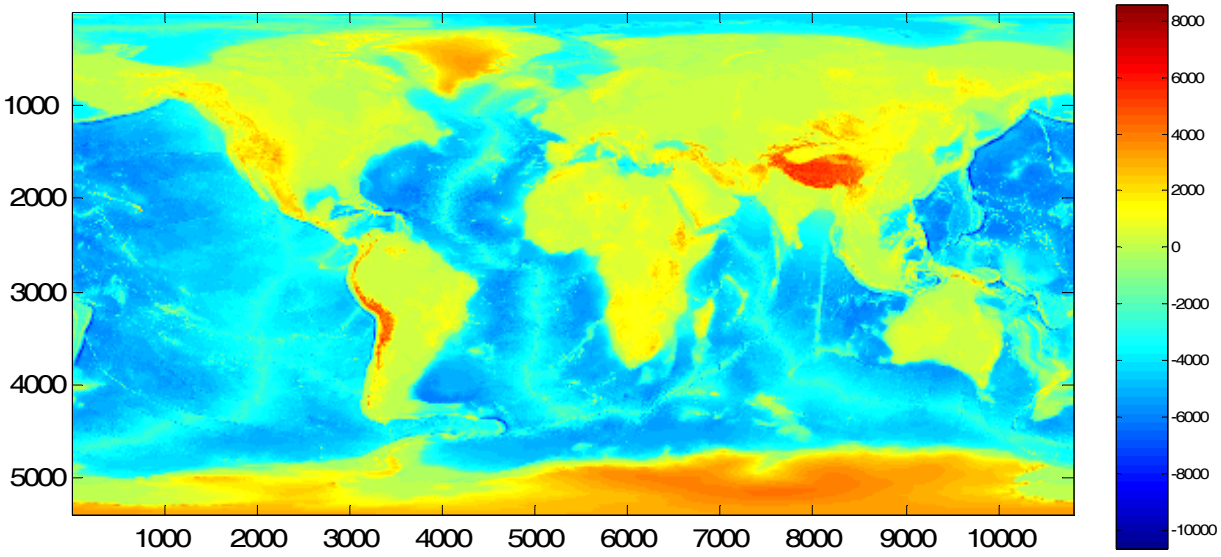


Figure 6: 2-Minute Gridded Global Relief Data (meters)

above sea level is known as topography and depth data below sea level is known as bathymetry; collectively, these data are known as relief. In the case of Earth relief, the two regions are the above water portion – the inside, and the underwater portion – the outside. Thus in Hawaii, an island that sticks up out of the water is the inside region. By convention, the SDF is negative inside and positive outside, counter to relief data. The space of signed distance functions is not linear because the sum of two signed distance functions is not a signed distance function [7]. In this section, low-dimensional manifolds will be determined using nonlinear dimensionality reduction by LLE for relief and for negated signed distance functions using data points consisting of small patches in coastal areas of the Earth.

The dataset used for Earth relief is the 2-Minute Gridded Global Relief Data [8], which has resolution of two minutes in both latitude and longitude. There are sixty minutes per degree and 360 total degrees of longitude on the earth, giving 10,800 columns of pixels. There are 180 total degrees of latitude and thus 5,400 rows in the dataset. Elevation is given in meters, as shown in Fig. 6. From this data, 2000 data points – 48 pixel × 48 pixel patches – were selected in the manner now described. A row and a column were selected uniformly and the 48 pixel × 48 pixel subimage with that row and column as its bottom left corner was examined. Patches not having at least 20% above water and at least 20% underwater pixels were rejected. The process was continued until 2000 points were accepted. The resulting data points are shown in Fig. 7, with

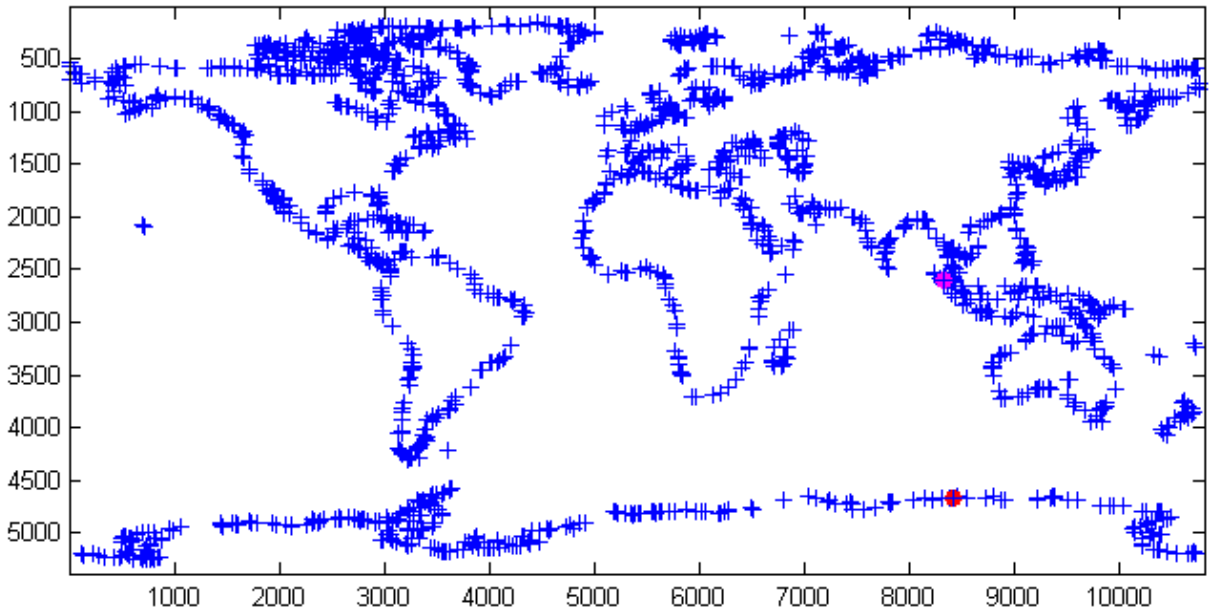


Figure 7: Southwest corners of 2,000 data points marked as +.

the crosses indicating the southwest corner of the patch. Examples of 48 × 48 patches are shown in the left column of Fig. 8. The top row is from a data point in Hawaii, the center row is from a data point in Antarctica highlighted with a red circle in Fig. 7, and the bottom row is from a data point in Southeast Asia highlighted with a magenta circle in Fig. 7. The relief data was treated as 48·48 = 2,304 dimensional for the purposes of LLE.

For each relief data point, an SDF representation was created as follows using a fast marching method [9]. In order to reduce boundary effects, first a larger patch of relief was taken for each

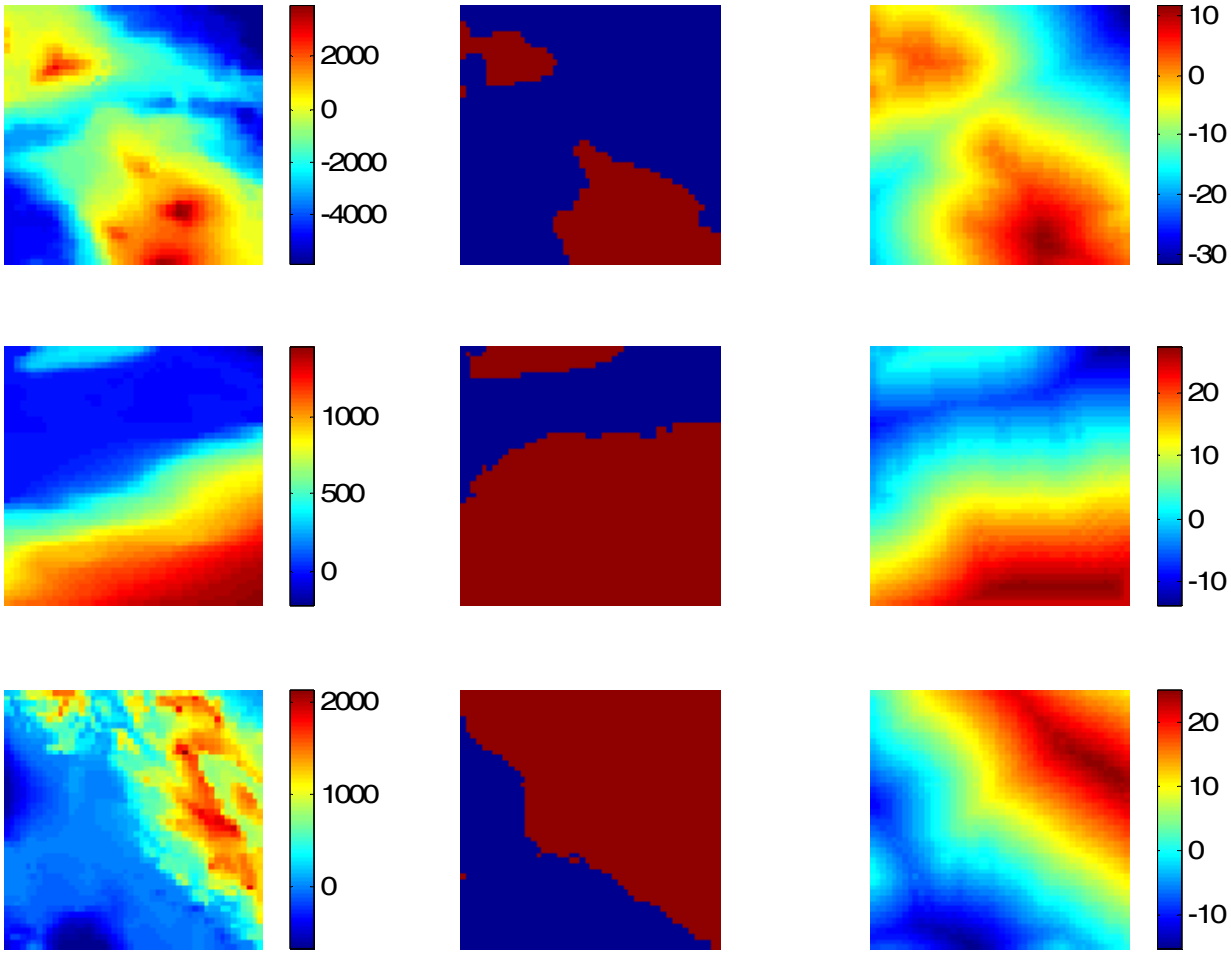
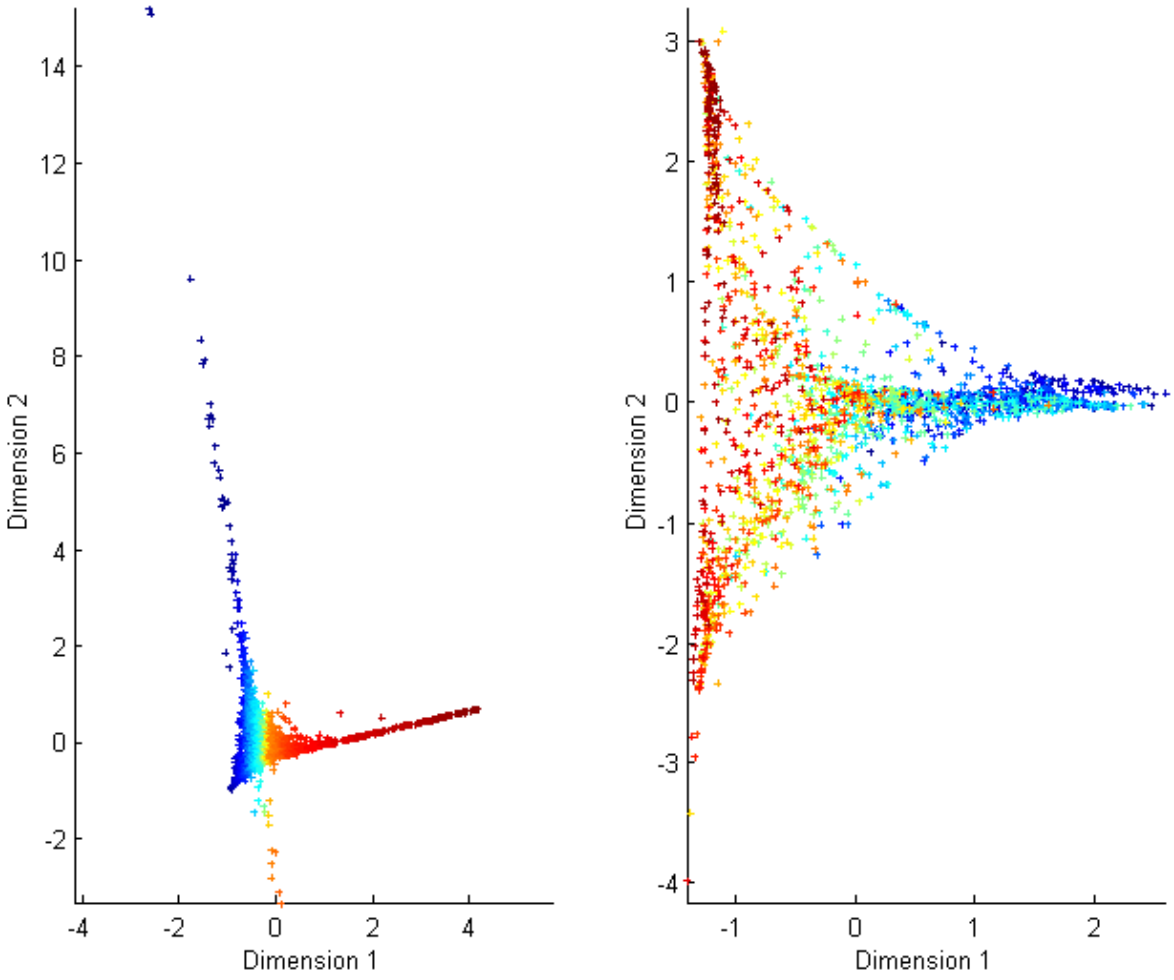


Figure 8: Relief data points (left column), binary image (center column), and SDF data points (right column).

data point. Specifically, 96 × 96 patches concentric with the 48 × 48 data points were used. For each larger patch, the relief was converted to a binary image demarcating the inside region and outside region. From the binary image, an SDF was generated using an implementation of the fast marching method [10]. Finally, the SDF image was clipped to be 48 × 48 and negated. Example 48 × 48 binary images and 48 × 48 negated signed distance functions are shown in the second and third columns of Fig. 8, respectively.

Comparing the true relief with the SDF, one can see that the two are not dissimilar, but there are differences. By definition, the SDF has constant slope, whereas in the true relief, the slope may be variable. Also, any ridges and valleys within a region that the true relief may have will not be captured by the SDF, as seen in the bottom row of Fig. 8. When the relief is primarily linear, then correspondence with the SDF is fairly good, as seen in the first two examples of Fig. 8.

Now, the low-dimensional manifolds will be compared as computed for $K = 6$. The first two dimensions of the relief data manifold are plotted in Fig. 9a and the first two dimensions of the SDF manifold are plotted in Fig. 9b as scatter plots. The colors in the relief data manifold are simply assigned linearly based on the dimension 1 coordinate. The color for each data point is transferred to the corresponding data point in the SDF manifold. It is apparent that there is a correlation between the two manifolds in the first dimension – as the first dimension coordinate



(a) (b)

Figure 9: Low-dimensional manifolds for (a) relief data and (b) SDF data.

in the relief manifold increases, the first dimension coordinate in the SDF manifold decreases. This inverse relationship is also apparent in Fig. 10, which plots the sorted dimension one coordinates of the relief manifold against the dimension one coordinates of the corresponding data points in the SDF manifold. Thus, despite certain features not being captured by the SDF representation, the primary mode of variation in both sets of data is similar.

The relationship in the second dimension of the low-dimensional manifolds is not as strong as in the first direction, but exists nonetheless. Fig. 11 shows this relationship in the same manner as Fig. 9. The distribution of colors in Fig. 11b, the SDF manifold, is certainly not random, but the pattern is more difficult to discern.

Another way to investigate the similarity of the two manifolds is to consider pairwise distances between the mappings of the data points. There may be metrics more or less appropriate for use in manifolds determined by LLE, but standard Euclidean distance was used in this case on the first three dimensions of each set. In the ideal case, pairwise distance in the relief manifold would be proportional to pairwise distance in the SDF manifold, so plotting them against each other would yield a straight line. Doing so with the actual pairwise distances, shown in Fig. 12

with the least-squares fit cubic function overlaid, does not give a straight line, but encouragingly, does give a monotonic trend. The conclusion from this analysis is that data points that are mapped close to each other in the relief manifold are also mapped close to each other in the SDF manifold.

An attempt will now be made to discover the underlying properties in the data that are the principal modes of variation. A variety of different local descriptors were investigated to determine whether they are

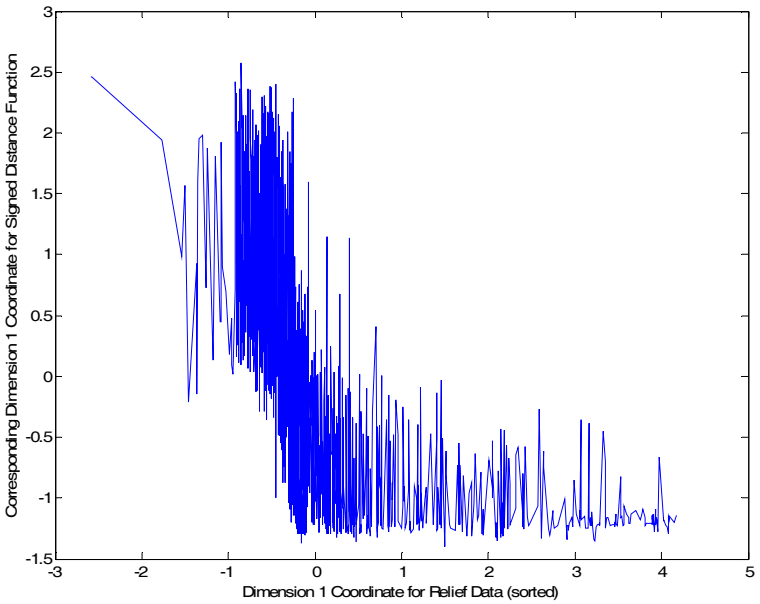
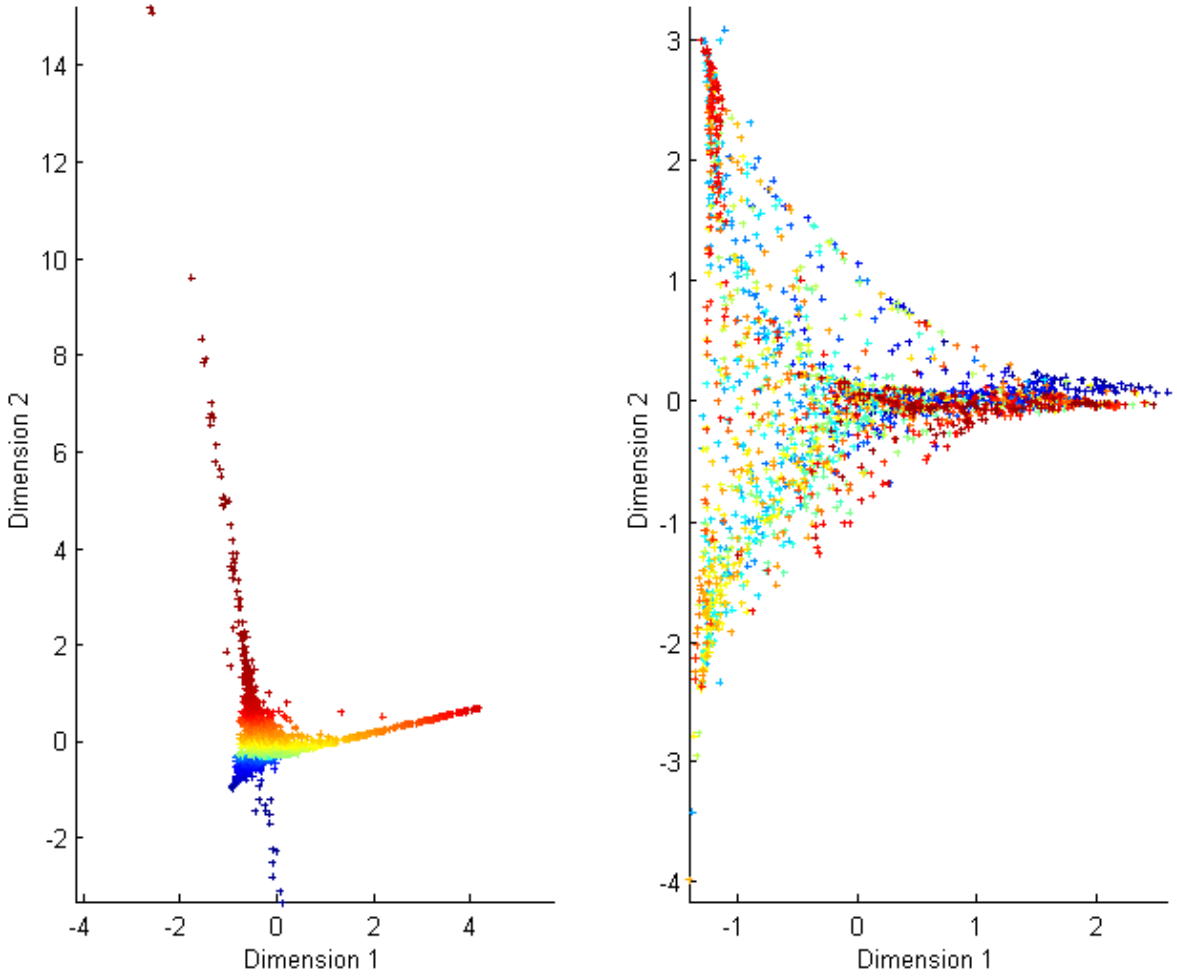


Figure 10: Correspondence between dimension 1 of relief and SDF low-dimensional manifolds.



(a) (b)

Figure 11: Low-dimensional manifolds for (a) relief data and (b) SDF data.

associated with the dimensions of the manifold. Two local descriptors were notable: the percentage of above water pixels in the image patch and the orientation of the gradients within the image patch.

As shown in Fig. 13, the second and third dimensions of the SDF manifold are extremely correlated to the proportion of pixels that are above water. The coloration in the figure shows the percentage of above water pixels in each data point. Red data points are 80% above water and blue data points are 20% above water. This relationship is very obvious from the figure. As demonstrated above, the connection between the SDF manifold and the relief manifold in dimension two exists but is not very strong. Consequently, the effect of the ratio of underwater pixels to above water pixels on the second and third dimension coordinates in the relief manifold is much less. As seen in Fig. 14, above water proportion is associated with how the data points get mapped to the manifold but it does not seem to be the principal effect.

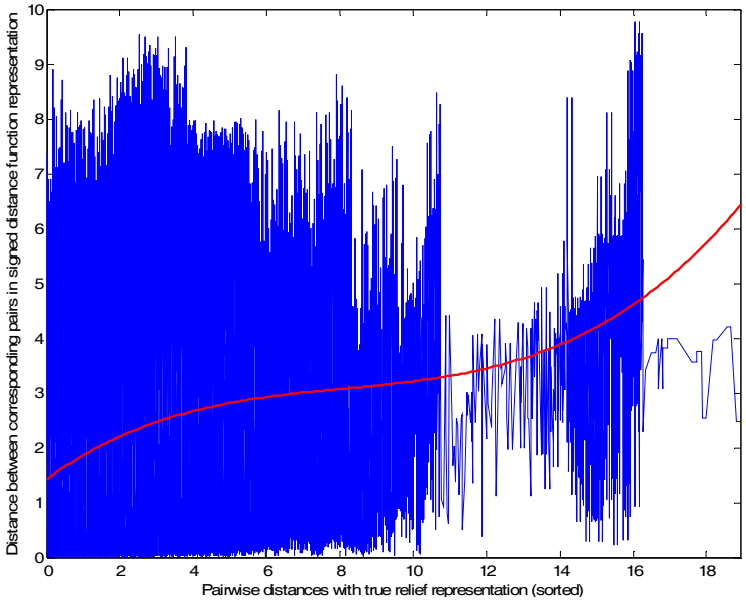


Figure 12: Correspondence between pairwise distances.

The second local descriptor found to have a strong relationship with the low-dimensional manifolds is the orientation of the gradients. A contour map for an example data point is shown for both the relief data and for the SDF in the first column of Fig. 15. The center column illustrates the gradient in an enlarged part of the image and the right column shows a histogram of the gradient directions with bins of width 30 degrees. The distribution of orientations in the relief image and the SDF image has the same general shape, but the distribution in the SDF image is peakier, due to the lack of ridges and valleys within a region that exist in relief. The full histogram provides a rich collection of local descriptors, but in the analysis below, the feature that is considered is the orientation bin with maximal count.

Fig. 16 demonstrates that the first dimension of the SDF manifold is related to the gradient orientation. The data points in the scatter plots are colored according to the maximum histogram bin. As dimension one in the SDF manifold increases, the maximum orientation completes a full cycle of angles counterclockwise. As the first dimensions of the SDF manifold and relief manifold are inversely proportional, the relief manifold's first dimension changes with gradient angle as well, but clockwise as the first dimension coordinate increases. This trend is shown in Fig. 17 using the same color scheme as Fig. 16.

Thus we have seen that two simple properties in the image patches can be used to label the axes in the low-dimensional manifolds. The signed distance function has many fewer degrees of

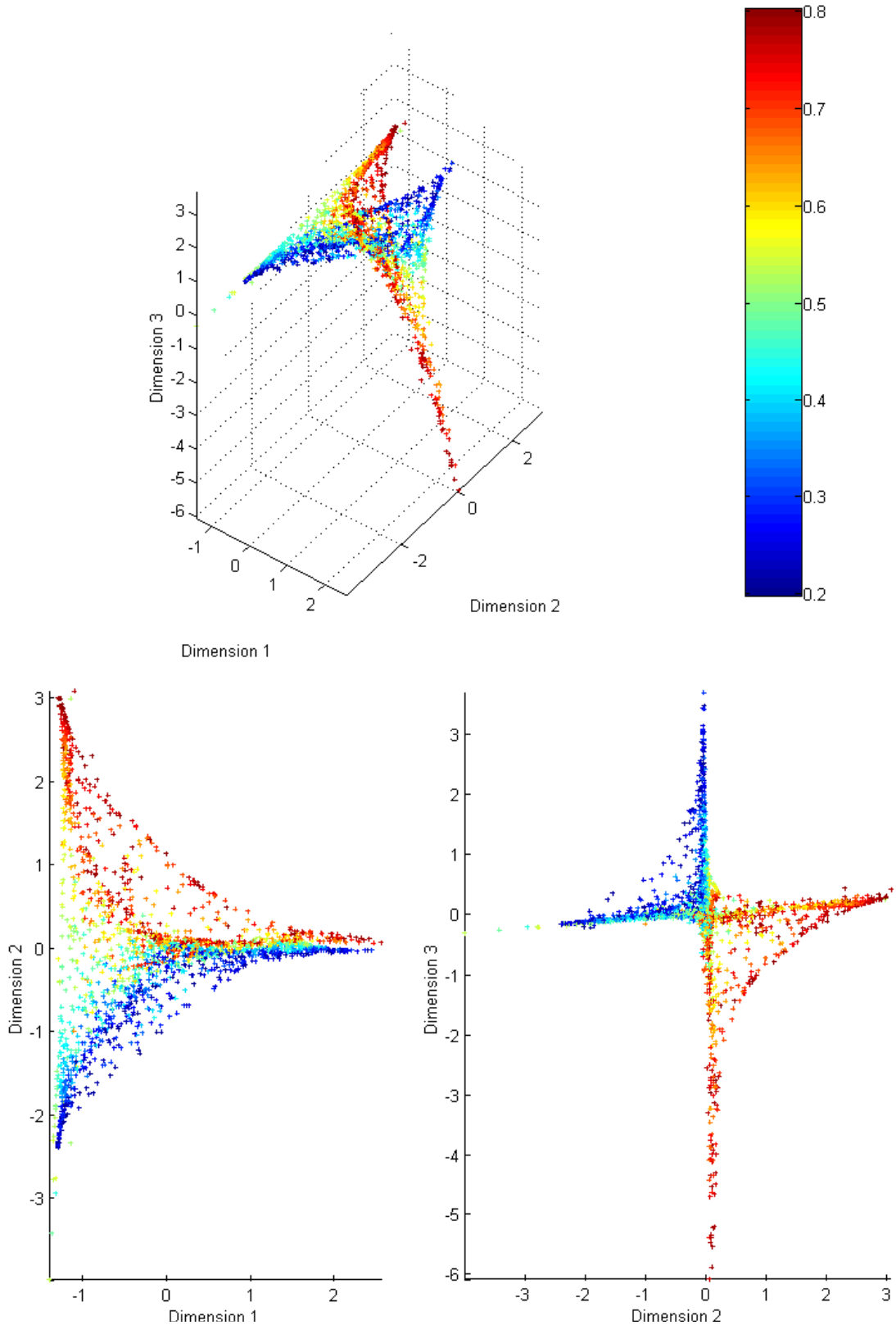


Figure 13: SDF manifold as a function of the proportion of above water pixels.

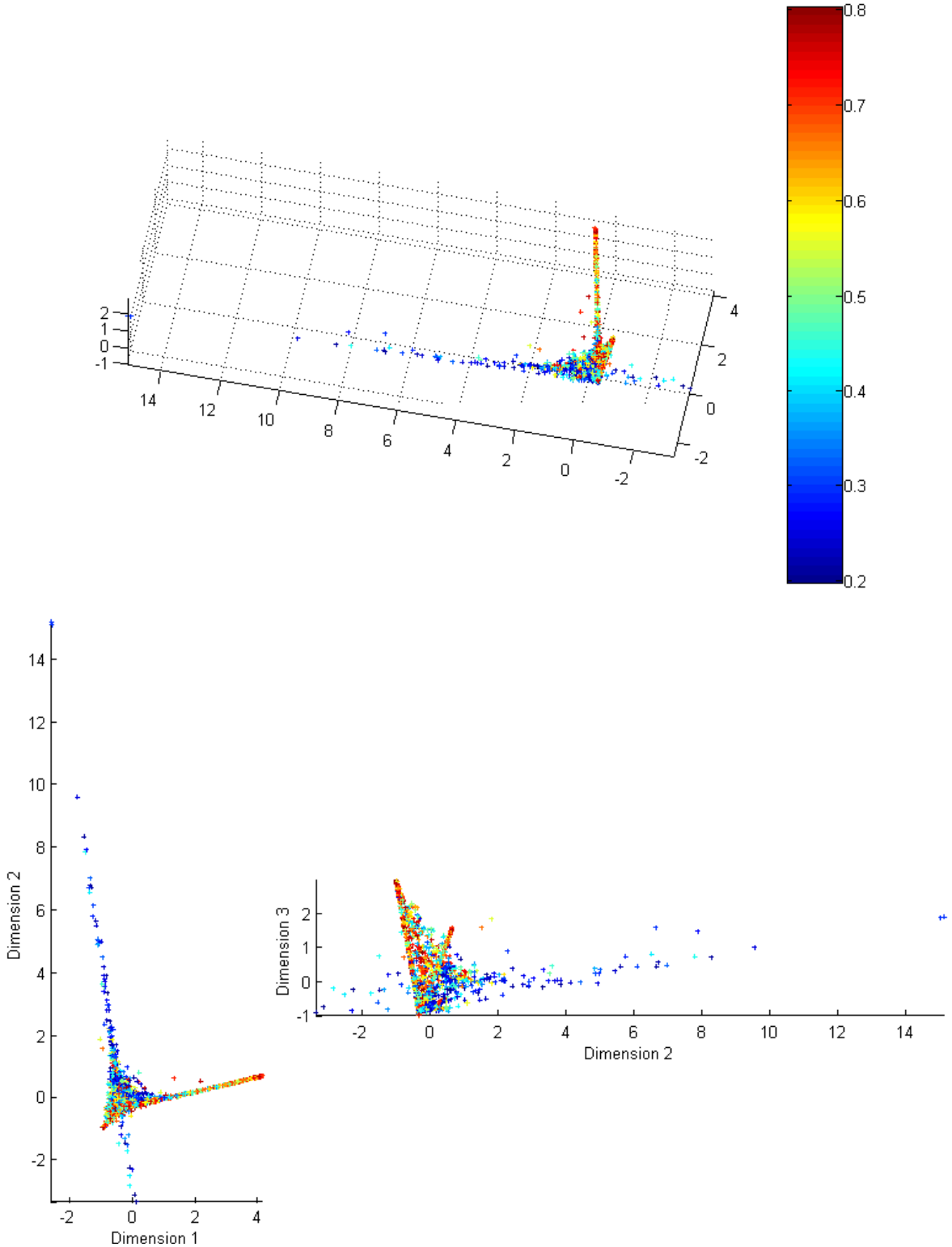


Figure 14: Relief manifold as a function of the proportion of above water pixels.

freedom than relief data, so the correlation with the maximum gradient angle and especially with the number of above water pixels is much stronger. However, despite the differences in representation, the same properties principally affect how the data points get mapped to lower

dimension. In some sense these properties are like scaling and rotation, and do not say much about the actual shape. That these two properties dominate is not surprising, but is counter to the intuition of what the space of shape should be. That is not to say that the actual form of the shape plays no role. Examining the histogram in ways other than looking at the maximal bin, for

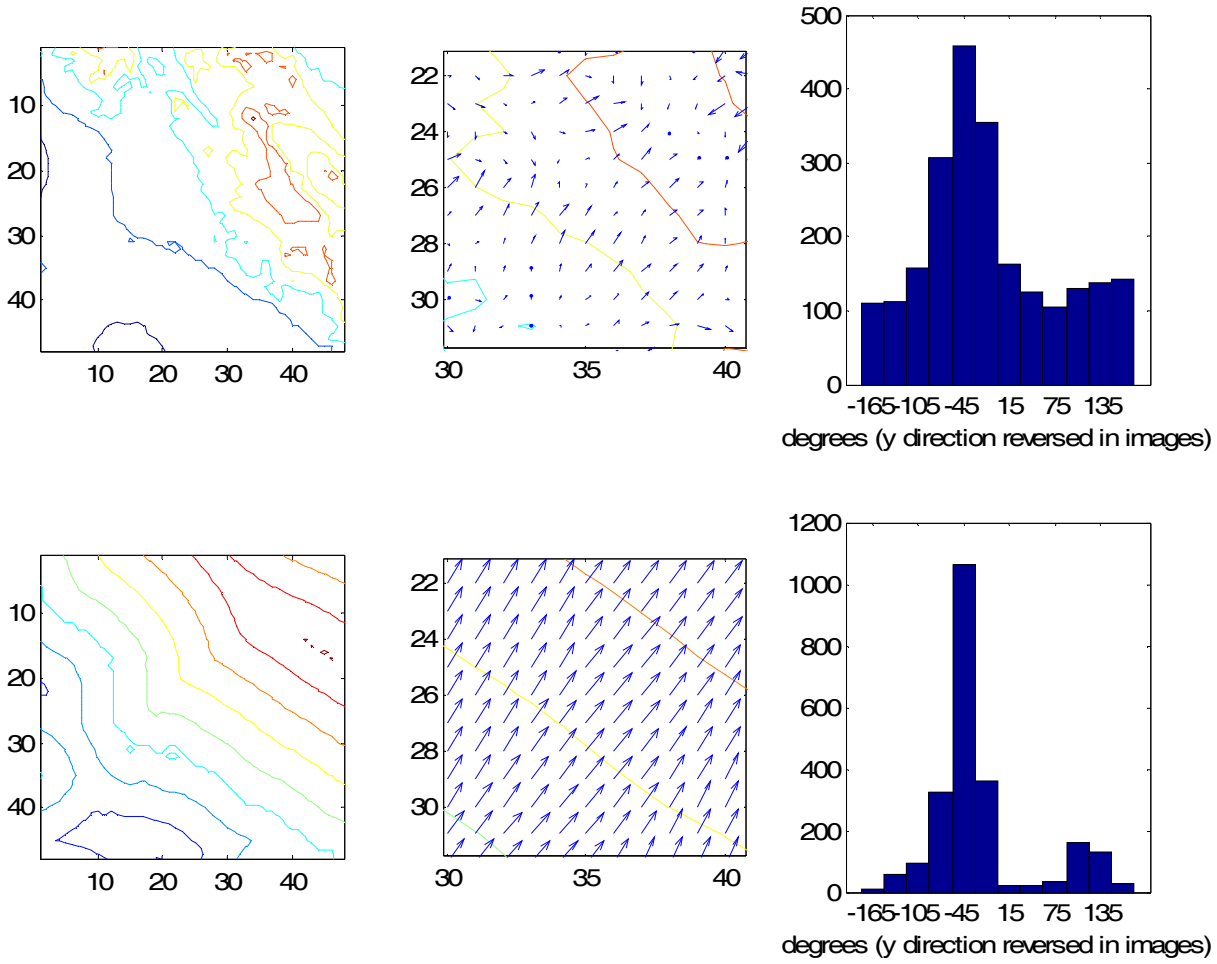


Figure 15: Examples of gradient angle histogram for relief data (top row) and SDF data (bottom row).

example, will give more insight to that role. In future work, it would be interesting to normalize the proportion of underwater pixels and to rotate all image patches to a common maximum angle before applying LLE. It may also be interesting to perform LLE on a scale invariant feature transform [11] of the image patches.

5. Conclusion

Nonlinear dimensionality reduction by locally linear embedding has been applied to social interaction data and to a space of signed distance function images along with relief images. Local and global structure in the low-dimensional space of social interactions coincides with what is expected intuitively. Additionally, modes of variation in the data are illuminated that would not have otherwise been seen. The dimensionality reduction approach seems like a

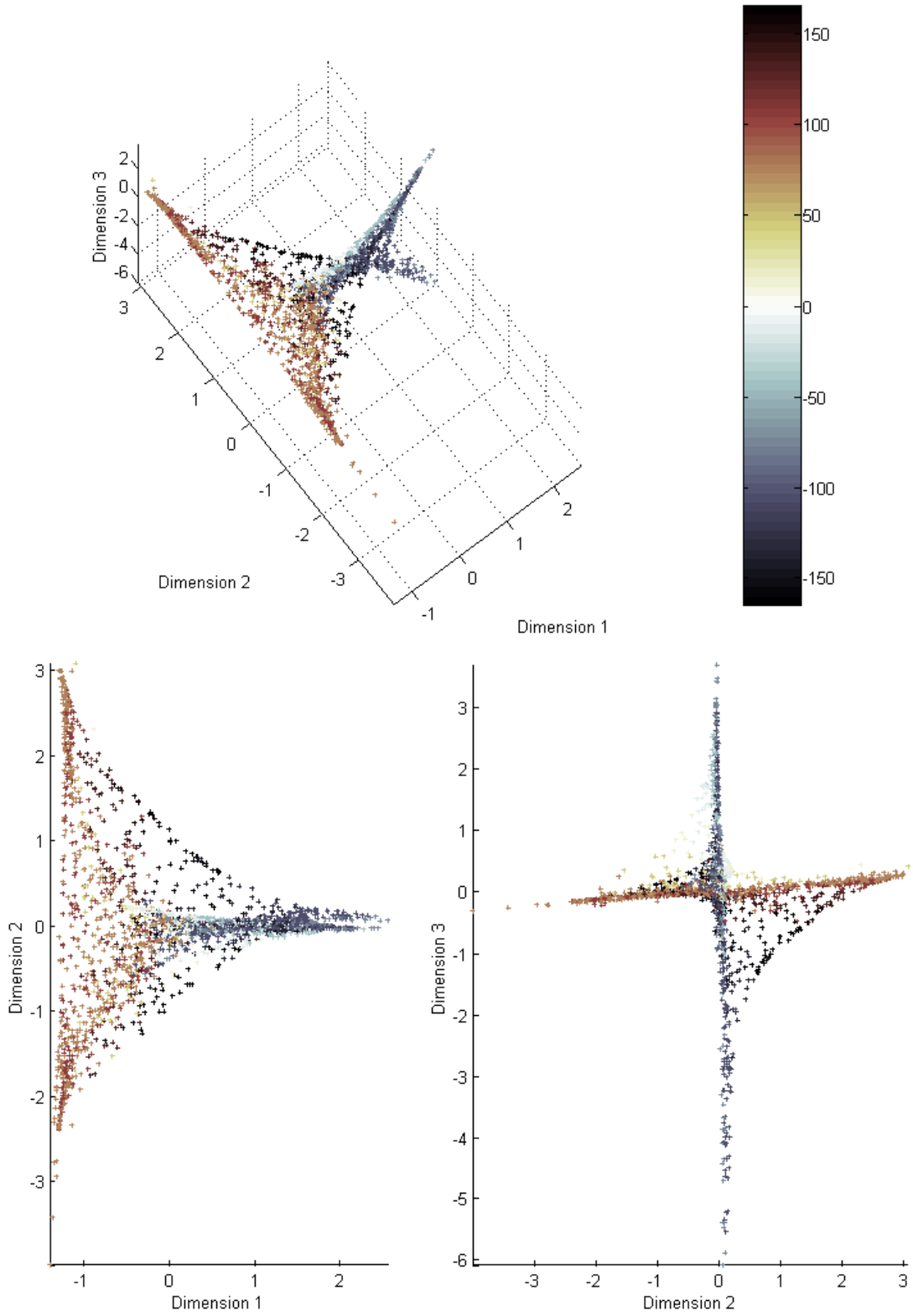


Figure 16: SDF manifold as a function of the maximum gradient angle.

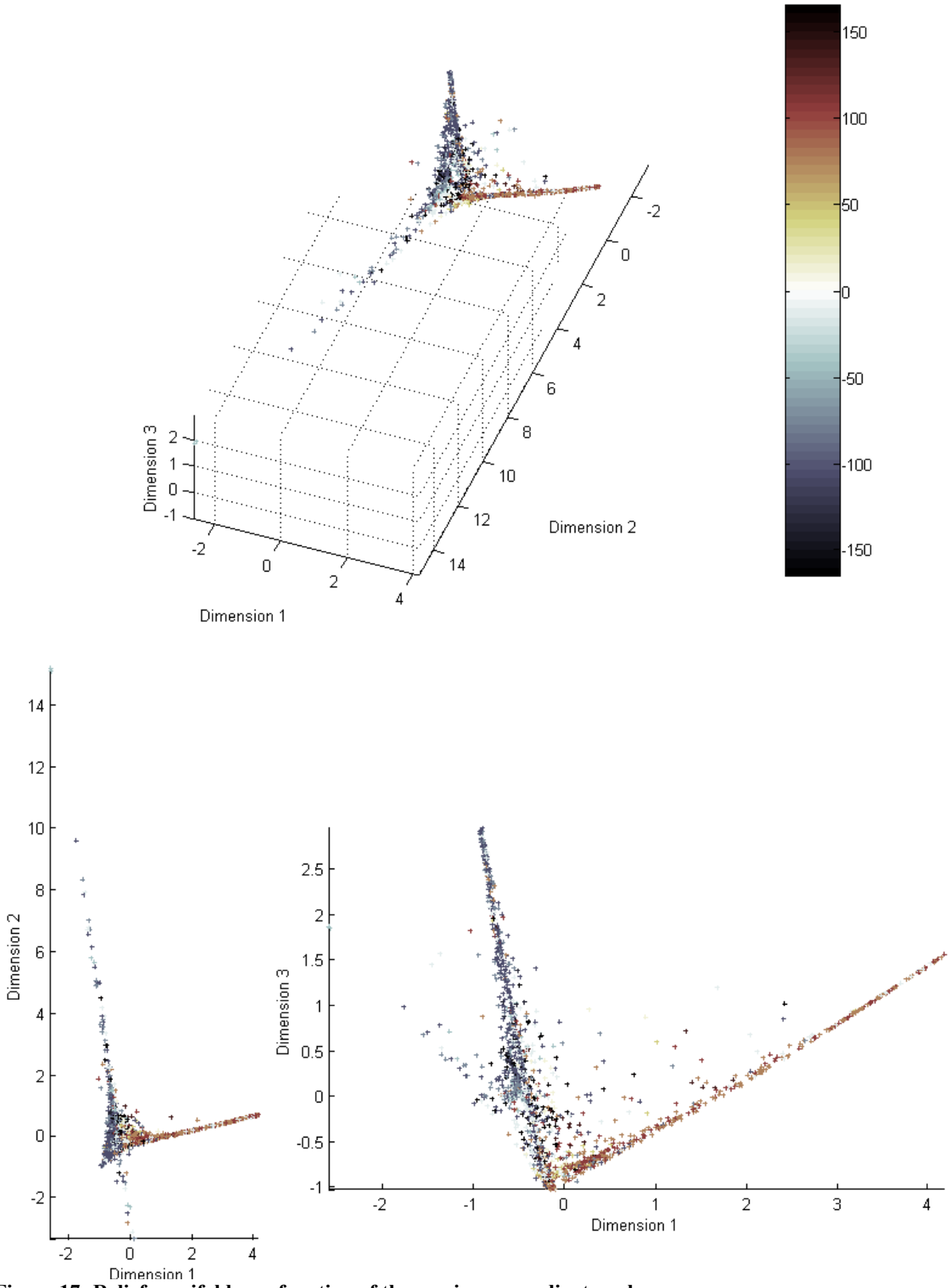


Figure 17: Relief manifold as a function of the maximum gradient angle.

promising way to construct or study social networks as an alternative to using graphs that connect two characters if they have an interaction [12]. Dimensionality reduction and specifically nonlinear dimensionality reduction by locally linear embedding is a versatile technique applicable to all forms of high-dimensional data analysis.

In comparing the low-dimensional manifolds produced by LLE for the relief data and the SDF data, it was shown that there is a good amount of correspondence. Thus, the analogy of equating the Hawaiian Islands to signed distance functions is fairly accurate. In addition, two simple local descriptors were found that are very closely related to the different axes in the low-dimensional manifolds. These two descriptors are tied to scaling and rotation; thus, it seems as though the structure of the space of signed distance functions is defined in large part by scaling and rotation, which are shape preserving, rather than by features more descriptive of shape. If manifolds arranged by shape properties are to be obtained through dimensionality reduction on the space of signed distance functions, it seems that scale and orientation will have to first be normalized.

References

[1] Sam T. Roweis and Lawrence K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323-2326, 22 Dec. 2000.

[2] Sam T. Roweis, 6.881 guest lecture, Cambridge, Massachusetts, Feb. 22, 2005.

[3] Lawrence K. Saul and Sam T. Roweis, "An introduction to locally linear embedding," unpublished. Available at: http://www.cs.toronto.edu/~roweis/lle/publications.html.

[4] Kisari Mohan Ganguli, The Mahabharata of Krishna-Dwaipayana Vyasa, Calcutta: Bharata Karyalaya Press, 1883-1896.

[5] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro, "Geodesic active contours," *International Journal of Computer Vision*, vol. 22, pp. 61-79, 1997.

[6] Alan S. Willsky, personal conversations, Sept. 2004 – present.

[7] Müjdat Çetin, Junmo Kim, and Alan S. Willsky, personal conversations, Sept. 2004 – present.

[8] National Geophysical Data Center, "2-minute gridded global relief data (ETOPO2)." Available at http://www.ngdc.noaa.gov/mgg/global/.

[9] James A. Sethian, "A fast marching level set method for monotonically advancing fronts," *Proceedings of the National Academy of Sciences*, vol. 93, pp. 1591-1595, 1996.

[10] Walter Sun and Venkat Chandar, "FMMgenDistMap."

[11] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91-110, 2004.

[12] Duncan J. Watts, Small Worlds: The Dynamics of Networks between Order and Randomness, Princeton: Princeton University Press, 1999.