# DUAL ROOTED-DIFFUSIONS FOR CLUSTERING AND CLASSIFICATION ON MANIFOLDS

*Steve Grikschat*[†], *Jose A. Costa*[‡], *Alfred O. Hero III*[†], *and Olivier Michel*[*]

[†]Dept. of Electrical Eng. and Computer Science, University of Michigan, Ann Arbor, MI 48109
[‡]Center for the Mathematics of Information, CalTech, Pasadena, CA 91125
[*]Laboratoire d'Astrophysique, University of Nice, 06108, France
Emails: {sgriksch, hero}@umich.edu, jcosta@caltech.edu, olivier.michel@unice.fr

## ABSTRACT

We introduce a new similarity measure between data points suited for clustering and classification on smooth manifolds. The proposed measure is constructed from a dual rooted graph diffusion over the feature vector space, obtained by growing dual rooted minimum spanning trees (MST) between data points. This diffusion model for pairwise affinities naturally accommodates the case where the feature distribution is supported on a lower dimensional manifold. When this affinity measure is combined with labeled data, a semi-supervised classifier can be defined that handles both labeled and unlabeled data in a seamless manner. We will illustrate our method for both simulated ground truth and real partially labeled data sets.

## 1. INTRODUCTION

Unsupervised learning or clustering have been a research focus of several communities for the past decades. Many algorithms have been proposed with varying degrees of success and are widely used in areas from text categorization/computer vision to genomics [1]. While many strides have been made in the area, there are still many open problems. Little success has been found in cases in which clusters do not form convex regions or are not clearly separated (overlapping). In particular, these scenarios can pose challenges to methods using Euclidean distances to measure similarities/affinities between data points. In a Euclidean space, a point on the edge of cluster 1 in Fig. 2 is closer to points in cluster 2 than to other points in cluster 1. In this case, there is no linear form which will classify the data satisfactorily.

Several different frameworks have been introduced to address the clustering problem. Classical solutions include generative model approaches and the $K$-means algorithm. In the generative model case, the data set is assumed to be well modeled by a parametric mixture of densities. One then attempts to estimate the mixture parameters via maximum likelihood or corresponding iterative implementations, such as the EM algorithm. Of course, model assumptions resulting in mathematical tractability and implementable inference algorithms will also result in a loss of capability of the method to cluster general data sets (that do not fit the assumed model). Another drawback stems from the usual lack of convexity of the likelihood function, resulting in an optimization problem which can have many local minima. The $K$-means algorithm, based on minimizing a mean squared error criteria (with respect to cluster centroids), shares with the previous approach the same problems in its optimization formulation.

Recently, the focus of attention in unsupervised learning has turned to spectral clustering methods due to its many successes [1]. These methods use the spectral content of a similarity matrix of pairwise distances between data samples to learn a partition of the data set. More specifically, the eigenvectors are viewed as providing an embedding of the data into a space where it is well separated and can easily be clustered.

While spectral methods have shown much improvement, there still remain difficulties. For example, starting from a set of measured dissimilarities, $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ (e.g., Euclidean distance), between pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ of data points, several promising spectral methods [2, 3] use a Gaussian or heat kernel to compute an affinity matrix, $\mathsf{A}$, according to:

$$a_{ij} = \mathrm{e}^{-\frac{d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)}{2\sigma^2}} \quad . \tag{1}$$

The kernel width parameter, $\sigma$ , gives the rate at which the similarity between two points decays. While there are many heuristic proposals for selecting the kernel parameter $\sigma$, there has been little effort to devise a systematic method for its determination. Complicating this matter, the direct reliance of spectral methods on the affinity matrix can cause clustering results to show high sensitivity to the choice of $\sigma$. This may lead to trial-and-error or other heuristic methods involving many re-starts for the selection of $\sigma$.

Despite their versatility, spectral methods still have trouble classifying data sets with non-Euclidean structures. For this we should seek out a more geometrically descriptive similarity measure: one that would better describe the global, as well as local, geometry of the data set. Recent research [4] have investigated diffusion processes on graphs. These processes are linked to a random walk on the graph with nodes consisting of data points. It is the isotropic growth on the graph which encapsulates the (assumed) underlying geometry of the feature space. In this paper, we provide a new affinity measured based on a similar graph diffusion idea. This framework naturally embodies a geometric point of view and is resistant to bottlenecks, noise, and non-convex/non-Euclidean structures. The method proposed is based on growing dual rooted minimum spanning trees (MST) between all pairs of points and using the hitting time of the two MST's to measure affinity between data points.

This research was supported in part by the National Science Foundation ITR grant No. CCR-0325571.

## 2. FROM DUAL ROOTED GRAPHS TO SIMILARITY MEASURES

Recent work by Coifman and Lafon [4] has provided a connection between diffusion processes on manifolds and random walks on finite data sets. Accordingly, a random walk on a finite data set can be seen as a discretization of a diffusion process, that will generate paths between data points with transition probabilities determined by local inter-point distances. By collecting all paths between any two points, one can naturally define a diffusion measure that accounts for the geometry of the data set: the more paths that connect two points the more similar they will be.

Motivated by this interpretation, we introduced a new similarity measure between data points based on the dual problem: starting two random walks on different points, when will the two paths generated hit each other? The following algorithm formalizes this idea. Let $\mathcal{X}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set of $n$ points in $\mathbb{R}^d$.

1. For each $\boldsymbol{x} \in \mathcal{X}_n$ compute a rooted greedy MST. This graph is obtained recursively in the following way. Let $MST_k(\boldsymbol{x}, \mathcal{X}_n)$ be the set of points in the tree at time $k$. Start with $MST_k(\boldsymbol{x}, \mathcal{X}_n) = \{\boldsymbol{x}\}$. Then, at time $k$, add the point $\boldsymbol{y}$ in $\mathcal{X}_n$ not previously added that is the closest (in Euclidean distance) to $MST_k(\boldsymbol{x}, \mathcal{X}_n)$, i.e.,

$$\boldsymbol{y} = \arg \min_{\boldsymbol{z} \in \mathcal{X}_n \setminus MST_k(\boldsymbol{x}, \mathcal{X}_n)} d(\boldsymbol{z}, MST_k(\boldsymbol{x}, \mathcal{X}_n)) \ .$$

2. Define the *hitting time*, $\tau(\boldsymbol{x}, \boldsymbol{y})$, between points $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{X}_n$ as the iteration $k$ when the two greedy MST's rooted on each point will intersect, i.e.,

$$\tau(\boldsymbol{x}, \boldsymbol{y}) = \min\{k : MST_k(\boldsymbol{x}, \mathcal{X}_n) \cap MST_k(\boldsymbol{y}, \mathcal{X}_n) \neq \emptyset\} \ . \tag{2}$$

3. The similarity/affinity between $\boldsymbol{x}$ and $\boldsymbol{y}$ is then determined according to the heat kernel (1) with $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tau(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Fig. 1 shows an example of a dual rooted tree obtained by running the above procedure in a two dimensional data set comprised of two clusters. Alternatively, one could also use the total length of the rooted trees at the hitting time to compute similarity between points. This measure has the appeal of accounting for the path lengths, as opposed to only counting the minimum number of steps to get from one point to the other.

Regarding implementation of the algorithm, $n$ full greedy MST's are grown and associated with a list containing the time stamps of when each data point is added to the each tree. To determine the hitting point, the ordered list of time stamps is parsed until a common point is found between trees.

## 3. APPLICATION TO SPECTRAL CLUSTERING

Although many flavors of spectral clustering have been proposed, they all share the same the algorithmic structure:

1. For a given affinity matrix $A$, define the diagonal matrix $D = \text{diag}(A\mathbf{1})$ and the graph Laplacian as $L = D - A$.
2. Solve the generalized eigenvalue problem

$$L\boldsymbol{v} = \lambda D\boldsymbol{v} \ .$$

3. Use the eigenvectors associated with the $k$ smallest positive eigenvalues to determine a $k$-way partitioning of the data. This will depend on the particular spectral clustering method chosen. It can range from heuristic based methods to applying $k$-means on the resulting eigenvectors [1, 3].

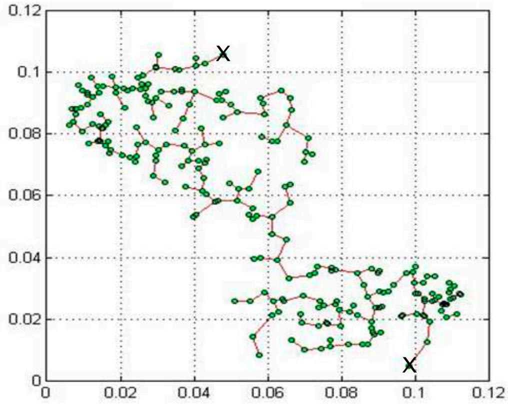

**Fig. 1**. Example of dual rooted MST. The '×'s' mark the two root vertices.

**Table 1**. Jaccard index for clustering of data sets from the UCI repository of machine learning [6].

| Data<br>set | NJW<br>+hit. time | NJW<br>+Euclid. | Ncut<br>+hit. time | Ncut<br>+Euclid. |
|---|---|---|---|---|
| **Wine** | .4047 | .4079 | .4047 | .4039 |
| **Wis. Cancer** | .4131 | .4231 | .4131 | .4397 |
| **Soybean** | .493 | .493 | .493 | .493 |
| **Housing** | .5508 | .5434 | .5508 | .5508 |
| **Ionosphere** | .4674 | .4261 | .4674 | .4289 |

### 3.1. Experimental Results

To test the algorithm, we implemented the proposed affinity measure with spectral algorithms introduced in [3] (NJW) and [2] (NCut).

To measure the efficiency of the clustering, we use a quantitative accuracy measure in addition to visual judgment. For this we use a measure known as the Jaccard index [5] between a predetermined set of class labels $C$ and a clustering result $K$:

$$J(C, K) = \frac{a}{a + b + c} \tag{3}$$

where $a$ is the number of pairs with the same class label in $C$ and the same cluster label in $K$, $b$ is the number of pairs with the same label in $C$ and a different label in $K$, and $c$ is the number of pairs with the same cluster in $K$ and different label in $C$.

Figures 2 and 3 show the results of applying the proposed method to some standard synthetic data sets. For comparison, we also applied $k$-means and spectral methods with an affinity matrix derived from Euclidean distances to the same data. From the figure, it is clear that $k$-means does not handle non-convex regions well. Also note that the spectral methods perform more accurately with the proposed dual rooted diffusion affinity matrix.

To illustrate the effectiveness of the proposed affinity measure in capturing the intrinsic geometry of manifold data, we applied the same algorithms to the "2 moons" data set embedded in a 3-dimensional space. As it can be seen from Fig. 4, the hitting time based affinity measure is clearly superior to the other methods.

Along with the synthetic 2− and 3−D datasets shown in figures 2, 3 and 4, the proposed method was also applied to standard real data sets of high dimension taken from the UCI repository of

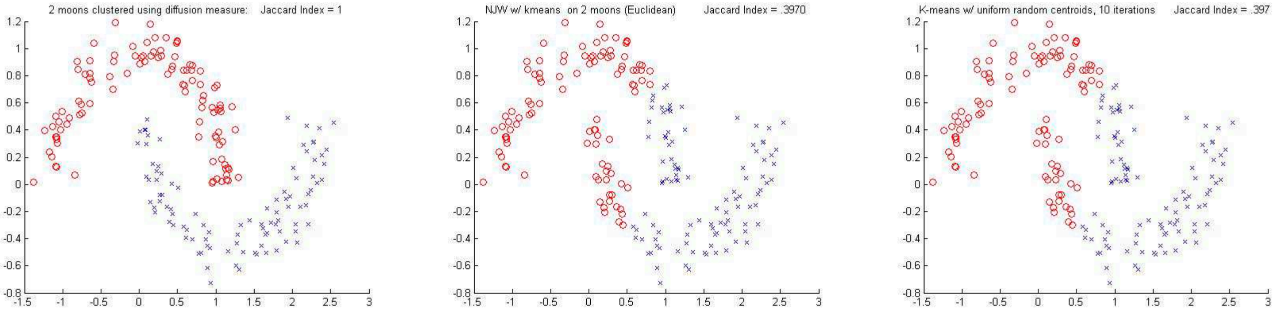

**Fig. 2**. Cluster results of "2 moons" data set. Cluster labels indicated by symbols. Left: NJW using hitting time affinity; Middle: NJW using Euclidean affinity; Right: $k$-means algorithm.

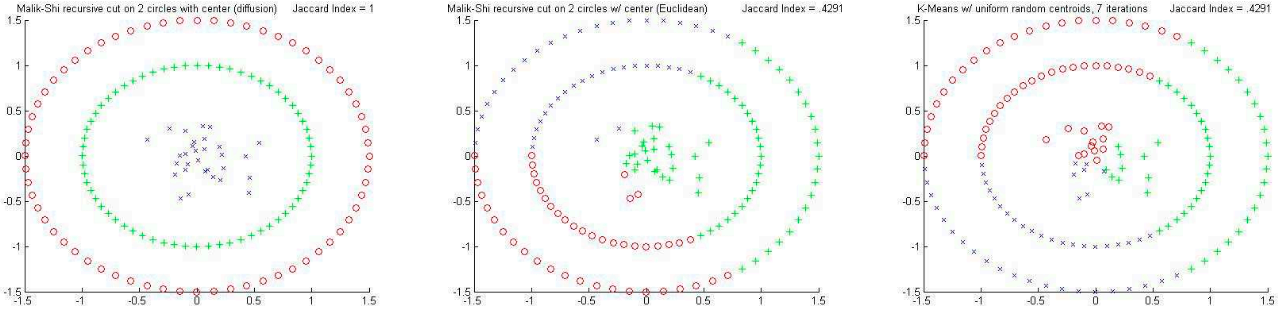

**Fig. 3**. Cluster results of "2 circles with center" data set. Cluster labels indicated by symbols. Left: NCut using hitting time affinity; Middle: NCut using Euclidean affinity; Right: $k$-means algorithm.

machine learning [6]. Table 1 shows the results of these experiments.

While for spectral methods based on Euclidean distances, the clustering solution is highly sensitive to the kernel width parameter $\sigma$, we have observed that the proposed affinity matrix yields solutions somewhat insensitive to this parameter. Given that $\tau(\boldsymbol{x}, \boldsymbol{y})$ is the number of steps in growing two rooted trees until they intersect, this implies that $\tau(\boldsymbol{x}, \boldsymbol{y}) \in \{1, \ldots, \lceil n/2 \rceil\}$. In practice, the parameter is typically chosen to be $10 - 50\%$ of this range. Fig. 5 shows this behavior.

## 4. APPLICATION TO SEMI-SUPERVISED LEARNING

Building on the recent work of Belkin et al [7], we can apply our similarity measure to the realm of semi-supervised learning. This type of feedback-based learning seems to be a more natural way of perform clustering, when extra label information is available. In [7], the intrinsic geometry of the feature space is factored into the formulation of functional learning through the use of the graph Laplacian in the extended optimization problem. Given $l$ data points $\{\boldsymbol{x}_i\}$ with labels $y_i \in \{-1, 1\}$ and $u$ unlabeled examples, let

$$f^* = \arg \min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(\boldsymbol{x}_i, y_i, f) + \gamma_a \|f\|_K^2 + \frac{\gamma_I}{u+l} f^T \boldsymbol{L} f \, ,$$

where $\mathcal{H}_K$ is a reproducing kernel Hilbert space, $V$ is a cost function used to to fit the labeled data, and $\| \cdot \|_K$ is the norm with respect to a kernel K. The application of the proposed affinity measure is apparent here: one can substitute our affinity matrix for the standard Euclidean based affinity in the computation of the graph Laplacian.

### 4.1. Experimental Results

We used the semi-supervised learning algorithms proposed in [7], namely regularized least squares (LapRLS) and support vector machines (LapSVM) with intrinsic geometric penalty via the graph Laplacian. For LapRLS, the cost function is $V(\boldsymbol{x}, y, f) = (y - f(\boldsymbol{x}))$, while for LapSVM the cost function is given by the hinge loss $V(\boldsymbol{x}, y, f) = (1 - y\, f(\boldsymbol{x}))_+$.

The effectiveness of the obtained labeling is measured according to the accuracy measure described in [8]. Given a set of true class labels $C$, the accuracy of a classification output $\bar{C}$ is defined as:

$$accuracy = \sum_{i>j} \frac{I\{I\{C_i = C_j\} = I\{\bar{C}_i = \bar{C}_j\}\}}{u(u-1)/2} \, , \qquad (4)$$

where $I\{\cdot\}$ is the indicator function.

Table 2 shows the results of applying the described methods to data sets with 10 to 20 labeled examples. As the choice of labeled

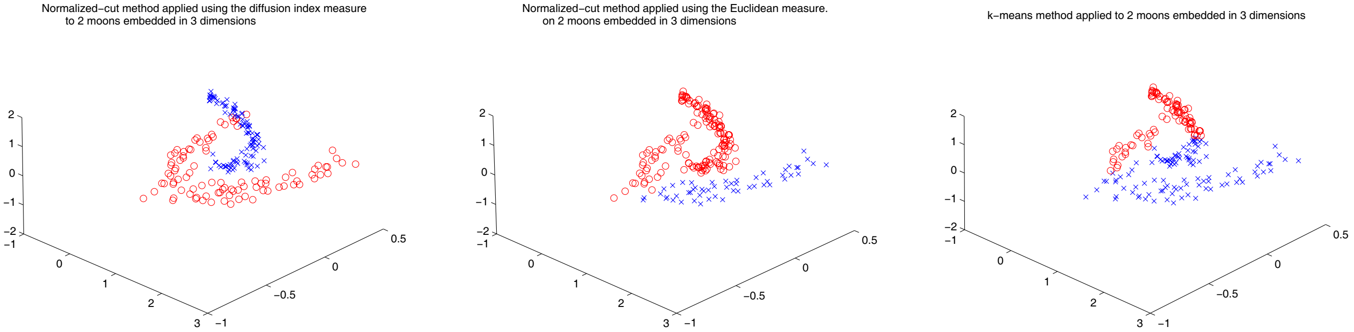

**Fig. 4**. Cluster results of "2 moons" embedded in 3-D. Cluster labels indicated by symbols. Left: NCut using hitting time affinity; Middle: NCut using Euclidean affinity; Right: $k$-means algorithm.

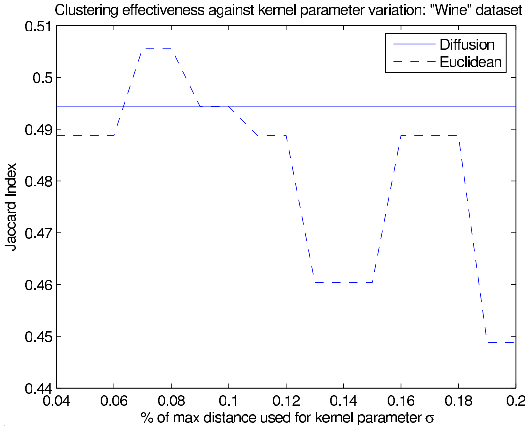

**Fig. 5**. Plot of variation in clustering effectiveness (Jaccard index) vs. change in heat kernel parameter $\sigma$. Experiment performed on Wine data set [6] using NCut method with both hitting time and Euclidean based affinities. Parameter chosen to be $\%$ of the maximum distance of data set.

examples can greatly affect the output, to avoid possible bias many simulations were run, with the labeled points chosen randomly for each experiment, but kept the same for different algorithms. The optimization parameters $\gamma_A$ and $\gamma_I$ were set by cross validating the accuracy measure. Typically, the best results were obtained when $\gamma_A \rightarrow \gamma_I$ or when the ambient space regularization was emphasized at least as much as the geometric regularization. As it can be observed, the use of the proposed hitting time affinity always outperforms the Euclidean based affinity. For some data sets, it even greatly improves the labeling accuracy.

## 5. CONCLUSION

We have introduced a new measure of dissimilarity between data points based on dual rooted diffusions. This results in a quantity that captures both local structure of the data set, through neighborhood relations, and global structure, through the complexity of the possible paths connecting any pair of points in the data set.

A natural extension of the ideas presented here is forming the affinity matrix using dual rooted graphs based on geodesic distances instead of Euclidean distances. This will allow one extra step in accounting for the intrinsic geometry of the data. We are also studying variations of the diffusion paradigm using single rooted trees. Of great importance to all this work is the theoretical characterization of the behavior of the proposed graphs.

**Table 2**. Accuracy of semi-supervised classification of data sets from the UCI repository of machine learning [6].

| Data<br>set | LapRLS<br>+hit. time | LapRLS<br>+Euclid. | LapSVM<br>+hit. time | LapSVM<br>+Euclid. |
|---|---|---|---|---|
| **Wis. Cancer** | .5463 | .5463 | .4071 | .4032 |
| **Housing** | **.8967** | .8108 | **.7641** | .7316 |
| **Ionosphere** | .5481 | .5461 | **.6278** | .5584 |

## 6. REFERENCES

[1] Y. Weiss, "Segmentation using eigenvectors: a unifying view," in *IEEE International Conference on Computer Vision*, 1999.

[2] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Analysis and Machine Inteligence*, vol. 22, no. 8, pp. 888–905, 2000.

[3] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Neural Information Processing Systems 14*, 2001.

[4] R. Coifman and S. Lafon, "Diffusion maps," *submitted to Applied and Computational Harmonic Analysis*, 2004.

[5] N. Bolshakova, "Cluster validity algorithms," `http://www.cs.tcd.ie/Nadia.Bolshakova/validation_algorithms.html`.

[6] S. Hettich, C. Blake, and C. Merz, "UCI repository of machine learning databases," `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[7] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: a geometric framework for learning from examples," Tech. Rep. 2004-06, Department of Computer Science, University of Chicago, 2004.

[8] E. Xing, A. Ng, M. Jordan, and S. Russel, "Distance metric learning, with application to clustering with side information," in *Neural Information Processing Systems 16*, 2003.