Multiscale Analysis on graphs via Diffusion

Mauro Maggioni
Mathematics and Computer Science
Duke University

DIMACS, 5/14/08

In collaboration with R.R. Coifman, P.W. Jones, Y-M. Jung, R. Schul, A.D. Szlam
Funding: ONR, NSF-DMS.
Plan

- Setting and Motivation
- Diffusion on Graphs
- Multiscale construction
- Conclusion
Structured data in high-dimensional spaces

A deluge of data: documents, web searching, customer databases, hyper-spectral imagery (satellite, biomedical, etc...), social networks, gene arrays, proteomics data, neurobiological signals, sensor networks, financial transactions, traffic statistics (automobilistic, computer networks)...

Common feature/assumption: data is given in a high dimensional space, however it has a much lower dimensional intrinsic geometry.

(i) physical constraints. For example the effective state-space of at least some proteins seems low-dimensional, at least when viewed at the time scale when important processes (e.g. folding) take place.

(ii) statistical constraints. For example many dependencies among word frequencies in a document corpus force the distribution of word frequency to low-dimensional, compared to the dimensionality of the whole space.
A deluge of data: documents, web searching, customer databases, hyper-spectral imagery (satellite, biomedical, etc...), social networks, gene arrays, proteomics data, neurobiological signals, sensor networks, financial transactions, traffic statistics (automobilistic, computer networks)...

Common feature/assumption: data is given in a high dimensional space, however it has a much lower dimensional intrinsic geometry.

(i) physical constraints. For example the effective state-space of at least some proteins seems low-dimensional, at least when viewed at the time scale when important processes (e.g. folding) take place.

(ii) statistical constraints. For example many dependencies among word frequencies in a document corpus force the distribution of word frequency to low-dimensional, compared to the dimensionality of the whole space.
1000 Science News articles, from 8 different categories. We compute about 1000 coordinates, $i$-th coordinate of document $d$ represents frequency in document $d$ of the $i$-th word in a fixed dictionary.
1000 Science News articles, from 8 different categories. We compute about 1000 coordinates, $i$-th coordinate of document $d$ represents frequency in document $d$ of the $i$-th word in a fixed dictionary.
Data base of about 60,000 $28 \times 28$ gray-scale pictures of handwritten digits, collected by USPS. Point cloud in $\mathbb{R}^{28^2}$. Goal: automatic recognition.

Set of 10,000 picture (28 by 28 pixels) of 10 handwritten digits. Color represents the label (digit) of each point.
A simple example from Molecular Dynamics

[Joint with C. Clementi]

The dynamics of a small protein (12 atoms, $H$ atoms removed) in a bath of water molecules is approximated by a Langevin system of stochastic equations $\dot{x} = -\nabla U(x) + \dot{w}$. The set of states of the protein is a noisy ($\dot{w}$) set of points in $\mathbb{R}^{36}$.

Left: representation of an alanine dipeptide molecule. Right: embedding of the set of configurations.
Assume the data $X = \{x_i\} \subset \mathbb{R}^n$. Assume we can assign local similarities via a kernel function $K(x_i, x_j) \geq 0$.

Simplest example: $K_{\sigma}(x_i, x_j) = e^{-\|x_i-x_j\|^2/\sigma}$.

Model the data as a weighted graph $(G, E, W)$: vertices represent data points, edges connect $x_i, x_j$ with weight $W_{ij} := K(x_i, x_j)$, when positive. Let $D_{ii} = \sum_j W_{ij}$ and

$$P = D^{-1}W, \quad T = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}, \quad \Delta = (I - T)^{\text{symm.}} \quad \Delta = (I - T)^{\text{normalizedLaplacian}}$$

Note 1: $K$ depends on the type of data.

Note 2: $K$ should be “local”, i.e. close to 0 for points not sufficiently close.
Assume the data $X = \{x_i\} \subset \mathbb{R}^n$. Assume we can assign local similarities via a kernel function $K(x_i, x_j) \geq 0$.

Simplest example: $K_\sigma(x_i, x_j) = e^{-||x_i-x_j||^2/\sigma}$.

Model the data as a weighted graph $(G, E, W)$: vertices represent data points, edges connect $x_i, x_j$ with weight $W_{ij} := K(x_i, x_j)$, when positive. Let $D_{ii} = \sum_j W_{ij}$ and

$$P = D^{-1}W, \quad T = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}, \quad \Delta = (I - T)$$

Note 1: $K$ depends on the type of data.

Note 2: $K$ should be “local”, i.e. close to 0 for points not sufficiently close.
Assume the data $X = \{x_i\} \subset \mathbb{R}^n$. Assume we can assign local similarities via a kernel function $K(x_i, x_j) \geq 0$.

Simplest example: $K_\sigma(x_i, x_j) = e^{-||x_i - x_j||^2/\sigma}$.

Model the data as a weighted graph $(G, E, W)$: vertices represent data points, edges connect $x_i, x_j$ with weight $W_{ij} := K(x_i, x_j)$, when positive. Let $D_{ii} = \sum_j W_{ij}$ and

$$P = D^{-1}W, \quad T = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}, \quad \Delta = (I - T)$$

Note 1: $K$ depends on the type of data.
Note 2: $K$ should be “local”, i.e. close to 0 for points not sufficiently close.
Let \( T \varphi_i = \lambda_i \varphi_i \) be the spectral decomposition of \( T \), with \( 1 = \lambda_0 \geq \lambda_1 \geq \ldots \), and \( \{ \varphi_i \} \) orthonormal bases for \( L^2(G) \). These eigenfunctions have been used in a variety of applications, in particular spectral partitioning, spectral embedding of graphs etc... Very little is in fact known about their properties.

All the pictures above were created by the map (called spectral embedding or eigenmap or diffusion map)

\[
G \rightarrow \mathbb{R}^d
\]

defined by

\[
\begin{align*}
    x &\mapsto (\sqrt{\lambda_i} \varphi_i(x))_{i=1,...,d} \\
\end{align*}
\]

For large \( t \),

\[
\| T^t \delta_x - T^t \delta_y \|_{L^2(G)} = \| (\lambda_i^{t/2} \varphi_i(x))_{i=0}^{\infty} - (\lambda_i^{t/2} \varphi_i(y))_{i=0}^{\infty} \|_{\ell^2(\mathbb{N})}.
\]

After truncation, we deduce the diffusion map embeds the graph in Euclidean space in such a way that this “diffusion distance” is approximately preserved.
Robustness with respect to nonlinear deformations

Mauro Maggioni
Multiscale Analysis on graphs via Diffusion
Robustness with respect to nonlinear deformations

Mauro Maggioni

Multiscale Analysis on graphs via Diffusion
We would like to be able to perform multiscale analysis of graphs, and of functions on graphs.

**Of:** produce coarser and coarser graphs, in some sense sketches of the original at different levels of resolution. This could allow a multiscale study of the geometry of graphs.

**On:** produce coarser and coarser functions on graphs, that allow, as wavelets do in low-dimensional Euclidean spaces, to analyse a function at different scales.

We tackle these two questions at once.
We construct multiscale analyses associated with a diffusion-like process $T$ on a space $X$, be it a manifold, a graph, or a point cloud. This gives:

(i) A coarsening of $X$ at different “geometric” scales, in a chain $X \rightarrow X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_j \cdots$;

(ii) A coarsening (or compression) of the process $T$ at all time scales $t_j = 2^j$, $\{ T_j = [ T^{2^j} ] \Phi_j \}_j$, each acting on the corresponding $X_j$;

(iii) A set of wavelet-like basis functions for analysis of functions (observables) on the manifold/graph/point cloud/set of states of the system.

All the above come with guarantees: the coarsened system $X_j$ and coarsened process $T_j$ have random walks “$\epsilon$-close” to $T^{2^j}$ on $X$. This comes at the cost of a very careful coarsening: up to $O(|X|^2)$ operations ($< O(|X|^3)$!), and only $O(|X|)$ in certain special classes of problems.
Multiscale Analysis, a sketch

Graphics by E. Monson
We now consider a simple example of a Markov chain on a graph with 8 states.

\[
T = \begin{pmatrix}
0.80 & 0.20 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.20 & 0.79 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.01 & 0.49 & 0.50 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.50 & 0.499 & 0.001 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.001 & 0.499 & 0.50 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.50 & 0.49 & 0.01 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.49 & 0.50 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.50 & 0.50 \\
\end{pmatrix}
\]

From the matrix it is clear that the states are grouped into four pairs \( \{\nu_1, \nu_2\}, \{\nu_3, \nu_4\}, \{\nu_5, \nu_6\}, \text{and} \{\nu_7, \nu_8\} \), with weak interactions between the pairs.
Some powers of the Markov chain $T$, $8 \times 8$, of decreasing effective rank.

Compressed representations $T_6 := T^{2^6}$ ($4 \times 4$), $T_{13} := T^{2^{13}}$ ($2 \times 2$), and corresponding soft clusters.
Multiscale Analysis, the spectral picture

Let \( T = D^{-\frac{1}{2}} WD^{-\frac{1}{2}} \) as above be the \( L^2 \)-normalized symmetric “random walk”.

The eigenvalues of \( T \) and its powers “typically” look like this:
Diagram for downsampling, orthogonalization and operator compression. (All triangles are $\epsilon$—commutative by construction)
Properties of Diffusion Wavelets

- Multiscale analysis and wavelet transform
- Compact support and estimates on support sizes (not as good as one really would like!);
- Vanishing moments (w.r.t. low-frequency eigenfunctions);
- Bounds on the sizes of the approximation spaces (depend on the spectrum of $T$, which in turn depends on geometry);
- Approximation and stability guarantees of the construction (tested in practice).

One can also construct diffusion wavelet packets, and therefore quickly-searchable libraries of waveforms.
Diffusion Wavelets on Dumbell manifold

Mauro Maggioni
Multiscale Analysis on graphs via Diffusion
Example: Multiscale text document organization

Mauro Maggioni
Multiscale Analysis on graphs via Diffusion
<table>
<thead>
<tr>
<th>$\phi_{2,3}$</th>
<th>Document Titles</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acid rain and agricultural pollution</td>
<td>nitrogen, plant, ecologist, carbon, global</td>
</tr>
<tr>
<td></td>
<td>Nitrogen’s Increasing Impact in agriculture</td>
<td></td>
</tr>
<tr>
<td>$\phi_{3,3}$</td>
<td>Racing the Waves Seismologists catch quakes Tsunami! At Lake Tahoe? How a middling quake made a giant tsunami Waves of Death Seabed slide blamed for deadly tsunami Earthquakes: The deadly side of geometry</td>
<td>earthquake, wave, fault, quake, tsunami</td>
</tr>
<tr>
<td>$\phi_{3,5}$</td>
<td>Hunting Prehistoric Hurricanes Extreme weather: Massive hurricanes Clearing the Air About Turbulence New map defines nation’s twister risk Southern twisters Oklahoma Tornado Sets Wind Record</td>
<td>tornado, storm, wind, tornadoe, speed</td>
</tr>
</tbody>
</table>

Some example of scaling functions on the documents, with some of the documents in their support, and some of the words most frequent in the documents.
Many open questions and applications

- How do properties of diffusion wavelets relate to geometric (multiscale) properties of graphs?
- How to visualize these multiscale decompositions?
- Better constructions?

Applied to
  - Multiscale signal processing (compression, denoising, discrimination) on graphs
  - Multiscale learning on graphs
  - Hierarchical clustering on nonlinear data sets.
  - ...
  - We will see at least a couple of applications to the analysis of networks and network traffic in other talks!
Acknowledgements

- R.R. Coifman, [Diffusion geometry; Diffusion wavelets; Uniformization via eigenfunctions; Multiscale Data Analysis], P.W. Jones (Yale Math), S.W. Zucker (Yale CS) [Diffusion geometry];
- P.W. Jones (Yale Math), R. Schul (UCLA) [Uniformization via eigenfunctions; nonhomogenous Brownian motion];
- S. Mahadevan (U.Mass CS) [Markov decision processes];
- A.D. Szlam (UCLA) [Diffusion wavelet packets, top-bottom multiscale analysis, linear and nonlinear image denoising, classification algorithms based on diffusion];
- G.L. Davis (Yale Pathology), R.R. Coifman, F.J. Warner (Yale Math), F.B. Geshwind , A. Coppi, R. DeVerse (Plain Sight Systems) [Hyperspectral Pathology];
- H. Mhaskar (Cal State, LA) [polynomial frames of diffusion wavelets];
- J.C. Bremer (Yale) [Diffusion wavelet packets, biorthogonal diffusion wavelets];
- M. Mahoney, P. Drineas (Yahoo Research) [Randomized algorithms for hyper-spectral imaging]
- J. Guinney, S. Lunagomez, J. Mattingly, S. Mukherjee, Q. Wu (Duke Math,Stat,ISDS) [stochastic systems and learning]; A. Lin, E. Monson (Duke Phys.) [Neuron-glia cell modeling]; D. Brady, R. Willett (Duke EE) [Compressed sensing and imaging]

Funding: ONR, NSF.

Thank you!

www.math.duke.edu/~mauro
Internet Multi-Resolution Analysis: Foundations, Applications and Practice


For more information:

www.ipam.ucla.edu/programs/mra2008/