Graph Diffusion & Math that Cuts

- We need to:

1. Build a graph
2. Wander around this graph
3. Decide which parts are should be clustered together and which points should be separated
4. Find and accept math to formalize & accomplish our goals

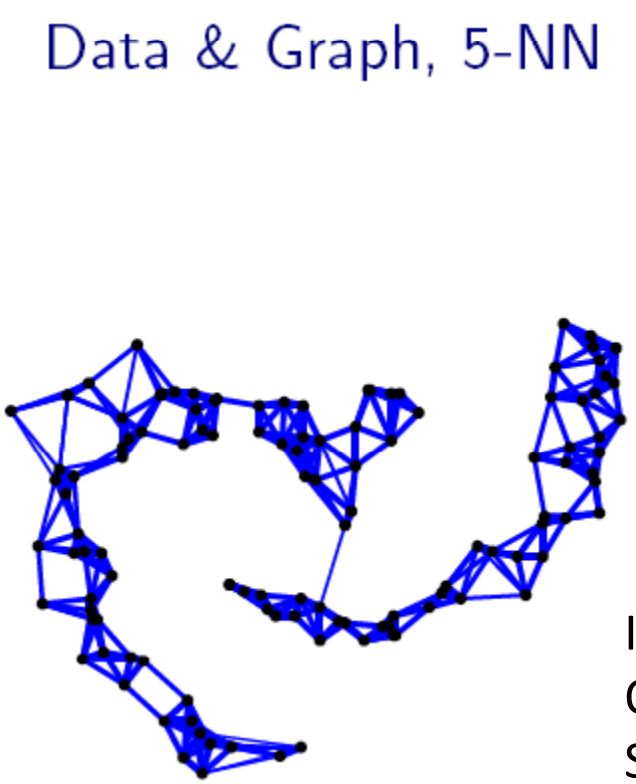
By David Jackson, 9/13/07

The Graph

- Tutorial on Spectral Clustering basically says: ‘a study which gives theoretical results about which graph one should use and when does not exit’
- Graph should be fully connected
- I think the following similarity function is reasonable as it is resistant to perturbation:

$$s(x_i, x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right)$$

The Graph 2:



The graph at left has a number of data Points that are connected based on Drawing a line from each point to it's 5 Nearest Neighbors. This graph method Is relatively easy to visualize.

Imagine defusing randomly around series Of connected nodes. Odds are you would Stay in either the curve on the left or the Curve on the right. The chances of "crossing" between the two curves appears small.

Core Idea:

- Want to cluster the data such that the wandering person spends as much time moving between points in the same cluster and as little time moving between points in different clusters as possible.

- How?

How? (A Proviso)

- Different people figured out ***that*** spectral worked before they fully figured out ***why*** it works.
- Various algorithms and explanations exist. Section 5-7 of the Technical Report describe the most prominent explanations.
- Hence early papers don't have the full story. Latter work is more involved. This weeks explanation will focus on intuition & accompanying math. A more rigorous story can follow.

How: The idea of Cut

- We want to express the chances of staying with in a cluster versus moving out of that cluster.
- From a node *i* on the graph we move randomly to any node connected to node *i.* Staying with in a cluster means that the probability of moving to a node outside of the cluster of *i* is low.

Moving on the Math

- For each point (i,j) the graph is defined as having a similarity S_{ij}. The matrix S_{ij} is the real valued adjacency matrix for the graph G.

- Let $d_i = \sum_{j \in List} S_{ij}$ be called the degree of node *i.*
- The volume of a set (cluster) $A \subset List$ be:

$$\mathrm{vol}\, A = \sum_{i \in A} d_i$$

NCut (a decent story)

$$NCut(A,\bar{A}) = \left(\frac{1}{vol\,A} + \frac{1}{vol\,\bar{A}}\right) \sum_{i \in A, j \in \bar{A}} S_{ij}$$

- $min_{assignments}(NCut)$ means finding subsets that we are unlikely to leave. We can use approximate methods to optimize (NP hard).
- Use a Laplacian matrix $L = D - S$ where D is a diagonal matrix form from the d_i. Solve:

$$Lx = \lambda Dx$$

[J. Shi and J. Malik, PAMI 2000]

A Random Walk View NCut

- Let's talk through this from another perspective
- Get the stochastic matrix by normalizing the similarity matrix S so all rows sum to 1:

- Define: $P = D^{-1}S$ (Solve: $Px = \lambda x$)

$$\pi_i^{\infty} = \frac{d_i}{vol\ List}$$

- With out much work you can show you have the stationary distribution of the markov chain. [Meila & Shi 2001]

Say “Abracadabra!”

- Let’s define: $P_{AB} = \Pr[A \to B|A]$

  As the prob. of moving from cluster A to B in one step of a random walk.

$$P_{AB} = \frac{\sum_{i \in A, j \in B} \pi_i^{\infty} P_{ij}}{\pi_i^{\infty}} = \frac{\sum_{i \in A, j \in B} S_{ij}}{vol(A)}$$

  Hence:

$$NCut(A, \bar{A}) = P_{A\bar{A}} + P_{\bar{A}A}$$