

Spectral Clustering on Similarity Based networks

s1235260

November 26, 2015

1 Abstract

The aim of this mini project is to study the evolution of interest based spacial communitites by exploiting the properties of the Laplacian matrix of a graph to carry out spectral clustering on communities with growing radii.

To test these methods we will use real data extracted from the popular dating app Tinder which allows us to collect nodes (people) within a certain radius in addition to biographies which we will use to create edges in between people to symbolize that they may know each other.

All together we will be using methods from machine learning (unsupervised), spectral graph theory and classical dynamics.

2 Physical Motivation

In this section I aim to motivate the use of eigen vectors in spectral community detection by taking of from interpreting the laplacian as the second derivative operator $(\nabla \cdot \nabla) = \nabla^2$:

First lets start of with a vector $\underline{u}(t)$ where each component $u_i(t)$ represents the amount of energy or heat in a node. Additionally we allow edges to permit heat flow in between nodes thus if one attempts to model the evolution of the system (of N nodes) with respect to determine the rate of $u_i(t)$ as the following:

$$\frac{d}{dt}u_i(t) = -\kappa \sum_{j=1}^N A_{ij}(u_i(t) - u_j(t))$$

Where :

$$A_{ij} = \begin{cases} 1 & \text{when } i \text{ is conected to } j \\ 0 & \text{otherwise} \end{cases}$$

Distributing the sum operator:

$$\frac{d}{dt}u_i(t) = -\kappa \left(u_i(t) \sum_{j=1}^N A_{ij} - \sum_{j=1}^N A_{ij}u_j(t) \right)$$

We know that if we sum up row 1 in matrix A we find out the number of nodes incident on node 1 which we define as the degree of node 1. Generlazing this to i follows trivially:

$$\frac{d}{dt}u_i(t) = -\kappa \left(u_i(t)deg(i) - \sum_{j=1}^N A_{ij}u_j(t) \right)$$

Since $A_{ii} = 0$ (A node connected to itself is trivial in our system and thus it is not accounted/allowed) we can use the kronecker delta symbol to rewrite this more compactly:

$$\delta_{ij} = \begin{cases} 1 & \text{when } i = j \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{d}{dt}u_i(t) = -\kappa \sum_{j=1}^N (\delta_{ij}deg(i) - A_{ij}u_j(t))$$

Which we can compact in to the vector matrix form (Where D is the diagonal degree matrix):

$$\frac{d}{dt}\underline{u}(t) = -\kappa \underbrace{(D - A)}_L \underline{u}(t)$$

This gives us a first order linear system of differential equations (similar to the heat equation if $L = \nabla^2$). Due to basic properties of matrix algebra solving this equation is rather trivial. I will provide a run through solution given that we ignore initial conditions for simplicity.

$$\frac{d}{dt}\underline{u}(t) = -\kappa L \underline{u}(t)$$

Lets guess the following ansatz to this ODE (common method in physics) as $\underline{v}e^{-\kappa\lambda t}$ where \underline{v} is some constant vector. Plugging in to the heat equation (and cancelling out equal terms) we now have:

$$\lambda \underline{v} = L \underline{v}$$

Which we recognize as a typical eigen value eigen vector equation. Which we can solve easily but the important thing I would like to comment on is what the solution means about our system. The different eigenvalues λ (also known as the spectrum of L) describe how every different solution (mode) to our heat equation changes dynamically since our solutions are of the form : $\underline{v}_i e^{-\kappa \lambda_i t}$, $i \in N$ so bigger lambdas may imply that our system is experiencing a decaying transfer of energy in time the eigen vector itself also tells us about the dynamics of the energy transfer since each of its component represents a node and in contrast to the others it tells us how the coupled system is simultaneously transferring energy (i.e. One node giving the other losing , energy transfer in synch etc..). Thus in a sense the eigen vectors by describing the different possible dynamics of energy transfer in the system tell us different things about connectivity in the graph and thus can help us find elements that may be clique like meaning that they all like to be connected to each other like a clique of friends.

3 Graph Construction

The first step is to construct a graph $G_{Tinder}(V, E)$ over which we will be using our spectral methods to study how communities evolve. In order to connect two people together we look at their biographies and based on a similarity measure which will be developed in this chapter we decide if they know each other or not.

3.1 How Tinder Works

Tinder is a mobile phone app which randomly pulls out people near by you (you define how near) and you swipe left or right depending if you like them or not. Most people include biographies and we also know the distance they are from you. So in this setting our phones are the origin of the ball of radius. Luckily for us we don't have to swipe since tinder has an API that allows us to pull out as many people as we want using a simple python wrapper.

3.2 Cosine Similarity Measure

The cosine similarity measure for two vectors \underline{v} and \underline{u} (s.t. $\underline{v}, \underline{u} \in \mathbb{R}^n$) is defined as the following:

$$\cos(\theta) = \frac{\underline{v} \cdot \underline{u}}{\|\underline{v}\| \|\underline{u}\|}$$

which just tells us how close two vectors are to each other. Now all we need to do this is transform our biographies into vectors.