Spectral Generation and Latent Community Structure of Multiweighted Networks*

Matthew Rocklin[1] and Ali Pinar[2]

[1]University of Chicago, Department of Computer Science, Chicago, IL
[2]Sandia National Laboratories, Livermore, CA†

Introduction Society uses networks to represent data either when linked interrelationships are natural (as is the case with hyperlinks and social networks) or when features are high-dimensional (such as large texts in document analysis.) Rather than sort such data as we would with numerical observations we can investigate community structure. Increasingly we find that our data has not one, but several interrelationships. Consider scientific journal articles - each may be linked by citations, text similarity, keyword similarity, shared authors, etc.... Each of these edge-types exhibit community structures that may be more or less correlated with each other. How do we cluster such multi-weighted data?

In an analog to principal component analysis (PCA) applied to multiple numerical observables we wish to find a few latent clusterings of this data which maximally explain the variation as compactly as possible. In this example we might find that articles can be independently classified by *topic* (Physics, Biology), *geographical region* (West Coast, East Coast, Midwest), and *institution type* (university, private industry, national lab). Such a compact description of each node in contrast to an otherwise incomprehensible table of network data is often useful in analysis.

To test our methods we develop a *spectral* network generation method which gives the ability to describe networks with these complex interrelated community structures. This method uses spectral theory to provide a transition from geometrically-embedded graphs (where structure is easy to describe) to non-embeddable graphs (which occur in nature) analogous to spectral graph clustering. This method is applicable to networks in general (not just multi-weighted) and we view it as one of our major contributions.

We present two contributions. First we propose a spectral network generation algorithm. Afterwards we discuss a method to draw out latent community structure from multiweighted graphs. We have encouraging results finding structure on spectrally generated multi-weighted networks.

Background and Related Work Networks generators exist [2, 3] which recognize the need for weighted networks with (overlapping) community structure and parameterized mixing. These systems work combinatorially, creating nodes and probabilistically connecting them with rules designed to obtain the desired effects. We found that these local methods were unsuitable for our needs of specifying complex large-scale community relationships. To solve this we devised a reversal of the spectral graph clustering method.

Spectral graph clustering[5, 6] is a community detection method that clusters a graph by finding a minimally disruptive embedding of the graph in low dimensions and then performing embedded graph clustering algorithms (such as k-means) on that data. It works by finding a low-rank approximation of the graph laplacian in hopes that the community structure signal will dominate and be preserved through the reduction. This low-rank approximation is performed by selecting those few eigenvectors with lowest eigenvalue. The

*This work is supported by the Laboratory Directed Research and Development program of Sandia National Laboratories.

†Sandia National Laboratories is a multi-program laboratory operated by Sandia Cor- poration, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. De- partment of Energys National Nuclear Security Administration under contract DE-AC04- 94AL85000.

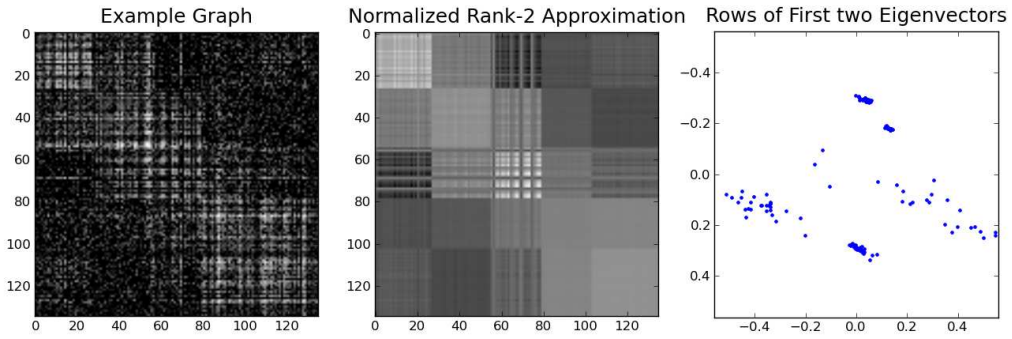

Figure 1: The Spectral Clustering process takes a full graph laplacian (first panel), finds a low-rank approximation of its normalized eigen-decomposition (second panel), and then considers the rows of the two eigenvectors as points in $\mathbb{R}^2$. We propose a method which traverses these panels *right to left*

rows of these vectors are considered as points in $\mathbb{R}^k$ where they should form obvious clusters, see Figure 1. Our proposed method is a reversal of this process.

To rebuild and "de-embed" this graph we must specify a supplemental eigenspace and create a spectrum. We can look towards spectral graph theory and spectra of real-world graphs to gain insight into this problem. Within this analysis we find parameters that affect important qualities of our graph.

Spectral Graph Generation

Spectral clustering takes a general graph and embeds it in $\mathbb{R}^k, k \ll n$ where intuitive clustering algorithms (such as k-means) can be effective. As humans we find this embedded space a more comfortable and intuitive place in which to work. However, we must not limit ourselves to only embedded graphs as natural networks are very unlikely to have this embeddable property. Spectral methods provide a transition between embedded graphs and general similarity-based networks. We take advantage of this transition in an algorithm (presented below) which is a near reversal of the steps found in [6]. For details please see this reference.

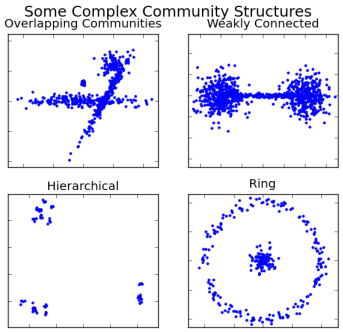

Figure 2: Community Structures encoded in 2D

Input: Points $y_i \in \mathbb{R}^k$ Representing community structure, Degree distribution, Spectral parameters

1. Form the points y_i into rows of an $n \times k+1$ matrix, U, prepending the ones vector.
2. Semi-normalize U by the degree distribution, D, $U \leftarrow D^{\frac{1}{2}}U$
3. Fill U randomly to be in $\mathbb{R}^{n \times n}$ and orthonormalize so that $U^T U = \mathbb{I}_n$
4. Create a spectrum (described below) $\Lambda_{ii} = \lambda_i$
5. Create the Symmetric, Normalized Graph Laplacian $L_{\text{sym}} = U\Lambda U^T$ and Laplacian $L = D^{\frac{1}{2}} L_{\text{sym}} D^{\frac{1}{2}}$
6. Return Similarity Weight matrix $W = D - L$

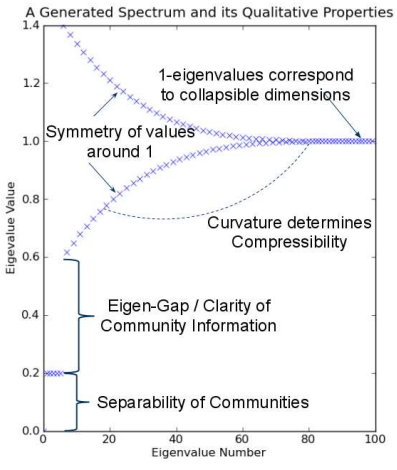

Figure 3: Example spectrum showcasing the values we use to parameterize graph spectra

That is, given a set of points embedded in $\mathbb{R}^k$, a degree distribution (optional), and a spectrum we can create a general, non-embeddable graph. The input set of points are arranged to encode community structure as in Figure 2. In this space it is simple and intuitive to describe complex relationships. The spectrum that we choose determines how this embedded graph is "de-embedded" or cast up to the space of general networks.

The spectrum with which we imbue our graph laplacian can affect how separated the communities are (algebraic connectivity), how clear they appear relative to the de-embedding information (eigen-gap), and how compressible or unembeddable the graph is (compressibility). For brevity, details on this topic are omitted but are available in [1] and are visually displayed in Figure 3. Additional insight may be gained from studying spectra of naturally occurring networks [1].

A Meta-Clustering Experiment

Spectral Graph Generation enables us to create multi-weighted graphs with a rich community structure. We perform an experiment which attempts to reconstruct latent community structure from several poor-quality graphs, each containing a very restricted view of that structure.

In Figure 4 we present three graphs embedded in 1-D, each having a distinct community structure which derives from a more complex 2-D embedded graph (the 3x3 grid). We can cast these 1-D graphs up to full networks using our spectral generation technique and can then cluster them to obtain different information from each. None of them individually gives the latent 3x3 structure.

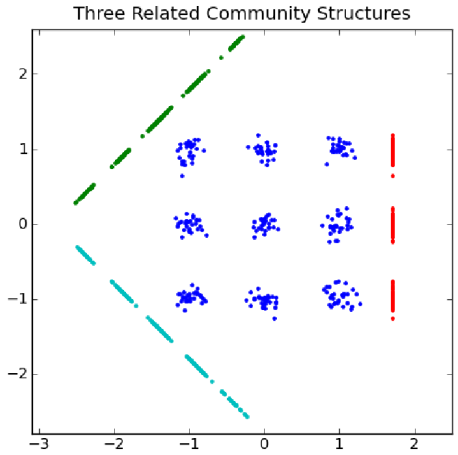

Figure 4: Three 1D graphs with community structures that all come from the same 2D source.

We perform the following experiment. Consider the above 3x3 embedded graph. Take many (we chose 16) 1-D random projections P_i of it. Cast each of those to a general graph $P_i \rightarrow G_i$ with spectral parameters which heavily obfuscate even the 1-D community structure (see Figure 5 - left). Consider many linear mixtures of these graphs $\mathcal{G}_j = \sum_i \alpha_i G_i : \alpha_i \in (-1, 1)$ and their clusterings $\mathcal{C}_j$ (we chose 1000 composite graph samples).

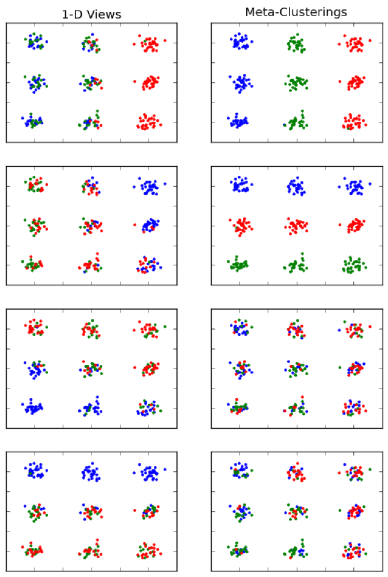

Figure 5: Clusterings from our experiment displayed on original 2d embedding. Left: clusterings of some input graphs. Right: Output clusterings, ordered by cumulative information content.

Represent this clustering data as its own graph with the clusterings as nodes and edge distances determined by the variation of information metric[4]. If many composite graphs have similar community structure then we can find *clusters of this graph of clusterings*. From each *cluster of clusterings* we compute an ensemble average[7] and are left with a few representative clusterings of our space of composite graphs which, as a set, encompass all the information of the underlying 3x3 grid.

We order these representative clusterings so that cumulative subsets maximize the set-wise information. That is, the first two representative-clusterings are maximally distant from each other and convey the most information as a set. The third is chosen to convey the next most information, conditional on the first two. This is an approximate solution to a set-generalization of Meila et. al's variation of information[4].

Note in Figure 5 how the first two representatives chosen are orthogonal in the information that they present. Knowing that a point is green in the first image gives no clue as to it's color in the second. Due to the ensemble averaging and multiple viewpoints they are also much cleaner than any of the original clusterings from the input graphs. We can recover the original 3x3 community structure using only the Cartesian product of the first two output clusterings.

References

[1] Anirban Banerjee. The Spectrum of the Graph Laplacian as a Tool for Analyzing Structure and Evolution of Networks. *University of Leipzig*, Ph.D. Thes, 2008.

[2] Andrea Lancichinetti and Santo Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 80(1):1–8, July 2009.

[3] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, March 2009.

[4] M Meila. Comparing clusteringsan information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, May 2007.

[5] Marina Meila and Jianbo Shi. A random walks view of spectral segmentation. *AI and Statistics (AISTATS)*, 2001:8–11, 2001.

[6] A Ng, M Jordan, and Y Weiss. On spectral clustering: Analysis and an algorithm. *Systems 14: Proceeding of the 2001*, 2001.

[7] Alexander Strehl and Joydeep Ghosh. Cluster EnsemblesA Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, 3(3):583–617, March 2003.