

ZMDS: Visualizing Structural Entropy in Queried Network Sub-Graphs

Justin Donaldson
Indiana University
1900 E.10th Street, PhD, 11th floor
Bloomington, IN 47406
+1 812-857-1160
jjdonald@indiana.edu

ABSTRACT

The large magnitudes of information in many scale free network datasets makes them difficult to analyze, plot, and visualize. Sacrifices often must be made on the plot's representational veracity or representational completeness. This paper discusses a method for exploring query based sub-graphs of a larger network using a method based on multidimensional scaling. Since the basis for the network data is a query, certain characteristics of node connections can be compared across the sub-graph and the original network. In this method, node weight data can be represented as a function of its global and local characteristics, exposing the negative entropy the node has in the given neighborhood. According to this scheme, a small sub graph of a larger network structure is analyzed as a modified and weighted Laplacian matrix. A z-score weighting scheme is used to modify each node's connection strengths in the neighborhood against its total number of connections in the original network, and the dimensionality of these weighted connection strengths is reduced to create a low dimensional embedding suitable for visualization and analysis. The motivation, methods, and example results are all discussed, and potential application methods are offered.

Categories and Subject Descriptors

H.1.1 [Systems and Information Theory]: Information Theory, Value of Information

H.3.1 [Content Analysis and Indexing]: Abstracting Methods

General Terms

Algorithms, Measurement, Experimentation

Keywords

Multidimensional scaling, structural entropy, MDS, network, query, visualization

1. INTRODUCTION

Laplacian matrices are a basis for representing network data as a matrix [1]. Several techniques, including Laplacian eigenmaps and spectral decomposition involve solving for low dimensional embeddings of network structure [2]. Usually, geodesic distance is used to encode connection weights, requiring that the matrix formatted network be positive semi-definite, or in network terms, symmetric. Other variants for network visualization include force-directed plots, Fruchterman-Reingold, and Kamada-Kawaii methods. However, these do not

involve a Laplacian basis for network representation, and so are not directly applicable to the techniques discussed in this paper.

Eigendecomposition methods produce a consistent representational form across any number of trials and orderings of data. This makes them ideal for machine learning and indexing techniques, such as the PageRank calculation used by Google [4]. However, the computation time and resources needed for large datasets of hundreds of thousands of nodes make this process intractable with conventional personal computing power.

In many cases, "querying" the network by extracting a significant collection of nodes and connections is a useful method of understanding more about local network structure. One such technique, called the "snowball" sampling method [5], involves selecting a collection of nodes and then expanding this selection with nodes with which they share a direct link. This method allows for an understanding of the original collection of nodes in the context of the connections they share with the larger network. A network constructed in such a fashion is called a "neighborhood" in this paper.

Unfortunately, scale free network characteristics of a graph will cause certain "hub" nodes to be included in query results at a much higher rate. In this context, hub nodes can constitute entropy, or non-salience in the plot representation. Even though they may share an above average number of connections in the queried neighborhood, a hub node's extra-neighborhood connections are often significantly higher than their local neighborhood connections.

2. Structural Entropy

The contrast between the neighborhood and the original network characteristics forms the basis for entropy in the structure of the network. Node weights are usually different between the two networks, and the magnitude of this difference is an indication of the entropy of that particular node in that particular neighborhood. Collections of nodes with negative entropy form network "structures" that are considered to have high negative entropy as well. The weights of connections between nodes indicate the strength of the connection according to one or more characteristics. In the following examples, playlist-based music data that exhibits scale free network characteristics will be used. According to this data, nodes are individual tracks (or songs), and the weights are the number of times these songs occur on a playlist. The neighborhood was constructed from a list of songs performed by several artists, including Jennifer Lopez, Bruce Springsteen, Tori Amos, Good Charlotte, and Oasis. An edge weighting scheme that preserves the variance (with a standard deviation unit form) is used. To achieve this, a modification to

the original Laplacian matrix definition is needed. First, a weighted adjacency matrix A is taken from the retrieved neighborhood. Next, we create the diagonal matrix D from the sum of the all intra and extra-neighborhood connection weights for each node in the neighborhood. Then, rather than subtracting the two matrices as in standard Laplacian form, they are added:

$$L_z = D + A$$

Finally, a weighting scheme is applied over L_z as a row or column operation:

$$w_{i,j} = \frac{k_{i,j} - \bar{k}}{\sigma}$$

With $w_{i,j}$ being the node weight between nodes i and j , $k_{i,j}$ being the co-occurrence counts between the nodes, and σ being the standard deviation for k . Applying this method for every k will make the matrix asymmetrical. However, this can be reconciled by applying a Euclidean distance calculation across L_z , using the same row or column orientation chosen in the previous step. The Euclidean distance matrix can be reduced to suitable dimensions using multidimensional scaling, and the resulting plot can be visualized, which produces an informative manifold using the previously mentioned music data.

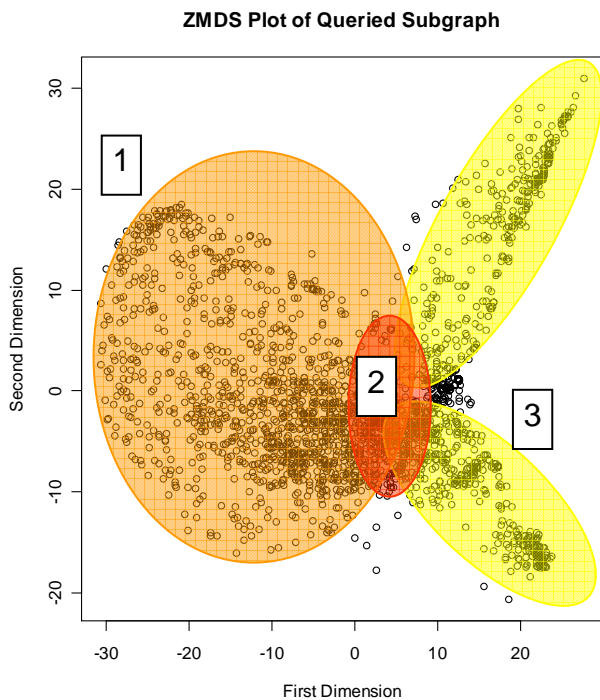


Figure 1. ZMDS representation of music playlist data

Two dimensional manifolds generated by this method often exhibit three noticeable components. A “Fan”, a “Zero Space”, and one or more “Tails”. Particularly noteworthy among these

components are the Tails (Feature 3 in Figure 1). Tails are evidence of high negative entropy in the structure of the neighborhood in question. They consist of clusters of nodes that form connections with themselves far more often than with other nodes in the context of the neighborhood. Since these connections are measured in terms of frequency, there is often a gradient of participation with the cluster. The nodes closest to the base of the Tail are like bridges from these tightly knit clusters to the rest of the neighborhood, while the nodes on the end of the tail only associate strongly with the cluster itself. In the plot in Figure 1, the Tails correspond to songs by Bruce Springsteen and Tori Amos, and the ZMDS representation shows that these two artists have songs that form connections with themselves far more often than with the rest of the neighborhood. The Tails also indicate which songs serve as “bridges” to the rest of the neighborhood. The base of the Tail usually attaches itself to a Zero Space (Feature 2 of Figure 1), where the entropy of the node structure passes zero. These nodes contain edge and identity weights close to zero as a result of the weighting function. This means that they are often hub nodes that are more specific to the neighborhood, participating very little with nodes from outside of the neighborhood. These nodes often connect the high entropy Tails to the larger Fan structure (Feature 1 in Figure 1). The Fan is a two dimensional representation of nodes that have more extra-neighborhood connections than intra-neighborhood connections, or that have smaller degrees and form the majority of their connections to nodes in the Fan.

3. APPLICATION AREAS

ZMDS has immediate applications in the realm of collaborative filtering and recommendation visualization. Other possible applications are currently under investigation.

4. ACKNOWLEDGEMENT

The author wishes to acknowledge MusicStrands, Inc. for their support of this research.

5. REFERENCES

- [1] Mohar, B. The laplacian spectrum of graphs. In Alavi, Chartrand, Ollermann and Schwenk (eds.). *Graph Theory Combinatorics and Applications*, Wiley, 871-898.
- [2] Belkin, M., and Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems 14*, Vancouver, British Columbia, Canada, 2002.
- [3] Fruchterman, T. M. J., and Reingold, E. M. Graph drawing by force-directed placement. *Software, Practice and Experience*, 21, 1129-1164.
- [4] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In Submitted to the 21st Annual ACM/SIGIR International Conference on Research.
- [5] Biernacki, P., Waldorf, D. Snowball sampling: Problems and techniques of chain referral sampling. *Sociol. Meth. Res.*, 10, 141-63.