Spectral Clustering of Attributed Graphs

Sahar Behzadi
University of Vienna
Vienna, Austria
sahar.behzadi@univie.ac.at

Claudia Plant
University of Vienna
Vienna, Austria
claudia.plant@univie.ac.at

ABSTRACT

To be complete

ACM Reference Format:
Sahar Behzadi and Claudia Plant. 2019. Spectral Clustering of Attributed Graphs. In *Proceedings of KDD 2019 (KDD'19).* ACM, New York, NY, USA, Article 4, **??** pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

TODO: Finish this draft intro Complex data in many domains can be represented as an attributed multidimensional networks. The nodes are characterized by often many attributes and links of potentially different types connect nodes. For instance in a social network the nodes represent users who are characterized by attributes like gender, age and hobbies. Users are connected by different types of links established e.g. by interaction on different platforms like Facebook and Twitter [?] TODO: This paper contains a dataset of twitter and facebook links with which we can experiment. In the analysis of neuroimaging data, anatomical brain regions are characterized by attributes like size and tissue type and are interconnected by structural and functional connectivity as measured by structural and functional magnetic resonance imaging [?]. TODO: Check if there is neuro data with this characteristics available

Clustering is very helpful in order to obtain an understanding of the major patterns in large attributed multidimensional networks. Ideally a cluster analysis should answer the following questions: Are there any dense subgraphs of a certain link type which are associated with a certain subspace of the attributes?

Notation

We consider an undirected attributed multidimensional network $G = (V, E, R)$. $V = \{v_1, ..., v_n\}$ represents the set of vertices. Each vertex v_i is characterized by c categorical attributes, where c is the dimensionality of the attribute space. We denote categorical attributes by upper case letters. A categorical attribute A has k distinct values, with $k = 2$ for binary attributes. We denote the value or category of vertex v_i in attribute A by $v_i.a$. We denote the total number of categories of all c categorical attributes by $|C|$. The set of edges E consists of quadruples $e = (v_i, v_j, w, r)$ where w represents a weight and r the type of the relation. We denote by $|R|$ the dimensionality of the network.

Problem Definition and Solution Overview

We consider the task of clustering and embedding multidimensional attributed networks. We address this challenge by defining a joint embedding of the mixed-type data to a low-dimensional space. In particular, we define a mapping $G \rightarrow \mathbb{R}^d$ minimizing the distance between connected vertices with similar attributes. When considering the setting of a single input graph without any node attributes as input data, our mapping is the same as Laplacian Eigenmaps [?]. When considering only categorical data without relational information, our mapping coincides with Homogeneity Analysis [?], a technique from statistics for categorical PCA. The low-dimensional coordinates obtained by our method are suitable for spectral clustering, i.e. subsequent application of K-means and for visualization.

Contributions

We propose spectralmix, a novel algorithm for spectral clustering and embedding of attributed multidimensional networks with the following properties:

- a comprehensive approach to dimensionality reduction and clustering of attributed multidimensional networks,
- is spectral embedding and -clustering when only applied to a single graph,
- is homogeneity analysis when applied to categorical data,
- integrates all available information when applied to multi-relational data with categorical node attributes,
- enables a sound interpretation of the clustering results.

2 A JOINT VECTOR SPACE EMBEDDING FOR MULTIDIMENSIONAL ATTRIBUTED GRAPHS

2.1 Objective Function

Our problem setting is characterized by mixed type data of different modalities, in particular we have a set of data objects which are characterized by multiple categorical attributes and linked by different relations or graphs. We integrate all the different modalities by a joint vector space representation. This joint vector space enables subsequent data mining. By application of the K-means algorithm we obtain a clustering inspired by spectral clustering methods. But we could also e.g., perform outlier detection or just visualize the data when we select a 2D or 3D vector space. In general, by embedding the data to a low-dimensional space we reduce noise and emphasize the major patterns. Our objective function for this embedding combines the ideas of spectral embedding of graphs with homogeneity analysis of categorical data. For n data objects we derive low-dimensional coordinates minimizing the following

KDD'19, August 2019, Anchorage, Alaska, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

objective function:

$$\min_{O,A} \sum_{i=i}^{|R|} \alpha_i \cdot \Big(\sum_{e \in E_i} w \cdot ||v_j - v_k|| \Big) + \sum_{i=1}^{|C|} \alpha_i \cdot ||v_j - a.j|| \quad \text{subject to the constraint that} C^T C = I_n. \tag{1}$$

$C \in \mathbb{R}^d$ is a $n \times d$ matrix of low-dimensional coordinates. Every row represents the coordinate vector for one data object. The first summand represents the contribution of the relational part of the data, i.e. of the $|R|$-dimensional graph on the position of a data object v_j. For every graph we consider all edges between v_j and vertices v_k which are neighbors of v_j and determine the coordinate matrix C such that the squared Euclidean distances are minimized. As in spectral embedding techniques we place the vertices as close as possible to their neighbors, but now not considering a single graph but multiple different relations. The second summand of the equation represents the categorical information. Categorical data establishes a bi-partite graph between data objects and their categories. Every categorical variable establishes its own bi-partite graph, where object v_j is connected to the category or value of its attribute $a(j)$. We now also consider these c categorical graphs in determining the joint vector space embedding. Inspired by Homogeneity Analysis, our algorithm Spectralmix will discover for every category a of every categorical attribute A a position in $\mathbb{R}^d$. In equation **??** we denote the position of the category of object v_j by $a.j$. The matrix $A \in |C| \times d$ contains these category coordinates as row vectors. TODO: find a more elegant notation if possible

The coefficients α_i represent weighting factors for the single modalities in the input data. It is of course possible to consider the unweighted objective function setting all α_i to 1. However, the total number of edges in the relation data and the categorical graphs tends to be very different as the categorical data establish bi-partite graphs where every node is connected to only and exactly one category. As default weighting scheme we therefore suggest to set the weight factors such that every modality has the same weight in the low-dimensional representation.

In order to avoid trivial solutions, we require the matrix C to be column-orthonormal. Without this constraint, a minimum of the objective function would be achieved by mapping all vertices to one common location.

2.2 Algorithm Spectralmix

TODO: describe algorithm based on pseudocode

2.3 Properties of Spectralmix

Here we need some proofs:

- The algorithm Spectralmix converges (to a global optimum of the objective function?)
- When considering only one single graph, the coordinates of the vertices are the same as in Laplacian Eigenmaps [?]
- Spectral clustering on these coordinates corresponds to the algorithm of Shi and Malik, see [?], this follows from the previous statement.

We also need to do a runtime complexity analysis.

```
algorithm Spectralmix $(G, C, d) \longrightarrow O \in \mathbb{R}^{n \times d}, A \in \mathbb{R}^{|C| \times d}$

//input: multidimensional graph $G$, categorical attribute matrix $C$,
desired dimensionality $d$
//output: $d$-dimensional feature space representation of data
objects $O$ and categories $A$; optional: clusters of data objects

    initialize matrices $O$ and $A$ randomly
    repeat
        //update object coordinates:
        $\forall i \in V : \mathbf{o}_i = \frac{1}{d_i} \sum_{i,j \in E} \mathbf{o}_j \in \mathbb{R}^d$;
        //update category coordinates:
        $\forall m \in C : \mathbf{c}_m = \frac{1}{n_m} \sum_{m \in E_c} o_m \in \mathbb{R}^d$;
        column-orthonormalize $C$;
        //e.g. with Gram-Schmidt to enforce $C^T C = I_n$
    until convergence;

    if clustering is desired
        perform K-means on the rows of $C$ with $k = d - 1$
        return clusters $c_1, ..., c_k$
    else
        return $O, C$;
```

Figure 1: The Algorithm Spectralmix.

3 INTERPRETATION OF THE RESULTS

3.1 Vector Space Embedding of Spectralmix: Proof of Concept

Spectralmix is a joint dimensionality reduction technique for multidimensional graphs with categorical node attributes. The results enable us to detect dependencies between graph structure and the categorical attributes. For a first proof of concept we consider the following synthetic data sets: Figure **??** displays the vector space embedding of a data represented by one unweighted undirected graph and one categorical attribute. The left panel in each subfigure uses the coordinates generated by SpectralMix for graph drawing and the right panel displays only the coordinates without the edges together with the category coordinates. Objects are colored according to the categorical attribute. Figure **??** (a) represents a synthetic data set where the graph and the categorical information agree. The graph consists of two clusters, where the first cluster is composed of objects having the category red and the second cluster of objects with blue category. There is some but not much noise in the data: In the graph, 90% of the links are between objects belonging to a common cluster and 10% between pairs of objects belonging to different clusters. Likewise, in the categorical data, 10% of the objects belong to the "wrong" category. It is evident that objects having the wrong category are mapped between both clusters. All important information is contained in the first generated coordinate.

See Figure **??**(b) for a data set with disagreement of both information. Categories have been assigned completely randomly with

50% red and blue. The first coordinate captures the graph information and the second coordinate the attribute information with is orthogonal to the graph. When using a random graph and a random attribute we have a different type of disagreement, see Figure **??**(c). Now, the embedding is driven by the categorical coordinate only, which is in the first dimension. The second orthogonal dimension is used for drawing the graph. See for comparison a drawing of only the graph information by Laplacian Eigenmaps where no structure is visible.

TODO:

3.2 Interpretation algorithm

Ideas:

- Quantify how much of the link information of each modality is preserved in each Eigenvector, i.e. find a score for what we see in the figures (this is easy...)
- Split up the modalities which provide orthogonal information in different groups
- Modalities might provide orthogonal information for parts of the objects and agree on different parts of the objects
- use information-theoretic concepts like mutual information, maybe MDL for parameterization

4 MIXED SPECTRAL CLUSTERING

We get cluster centers with K-means. This is easy but there is much open regarding the theoretical justification and the interpretation of these clusters.

TODO:

4.1 Graph Theoretical Analysis

Ideas:

- Search for work on spectral clustering multidimensional graphs
- Spectral clustering has a strong graph theoretical justification as relaxation of some kind of normalized graph cut (depending on the normalization of the graph Laplacian). How is it in our case? Can we give any justification of Spectralmix from a graph theoretic perspective? So, we have a multigraph plus $|C|$ bipartite graphs for the categories and what cut do we optimize?

4.2 Interpretation of the Clustering Result

K-means generates an additional categorical variable. We can do several things with that:

- study how cluster specific the categories are (this is easy)
- quantify how much the category locations and the graphs forces agree with the clustering (a bit harder)
- integrate K-means into the embedding procedure

5 EXPERIMENTS

TODO: This will be time-consuming; not so easy as with UCI data. Some data resources to be checked (but also look in related papers):

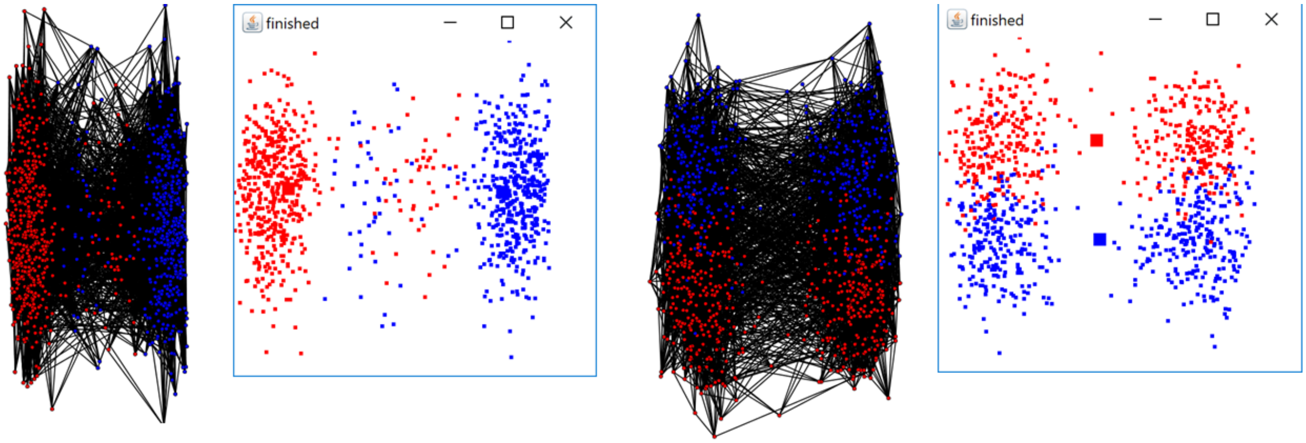
- friendship networks and sensor data [**?**], I already experimented a bit with this data, more data is maybe available at http://www.sociopatterns.org, see Figure **??**
- some attributed graphs are available here https://linqs.soe.ucsc.edu/data
- here are maybe some geographical networks which can be better embedded when we add categorical information about the location https://sparse.tamu.edu/about
- this website might also have some attributed graphs http://snap.stanford.edu/data/

To support very large networks a C implementation would be beneficial.

6 RELATED WORK AND DISCUSSION

CAMIR [**?**], HASCOP [**?**] and CLAMP [**?**]. Many of existing methods such as PICS [**?**],NNM [**?**], SSCG [**?**], UNCut [**?**], SA-Cluster [**?**], its faster version Inc-Cluster [**?**] and BAGC [**?**] perform hard clustering instead of allowing multiple memberships, i.e. fuzzy clustering that identifies overlapping clusters in which an object belongs with a membership probability.

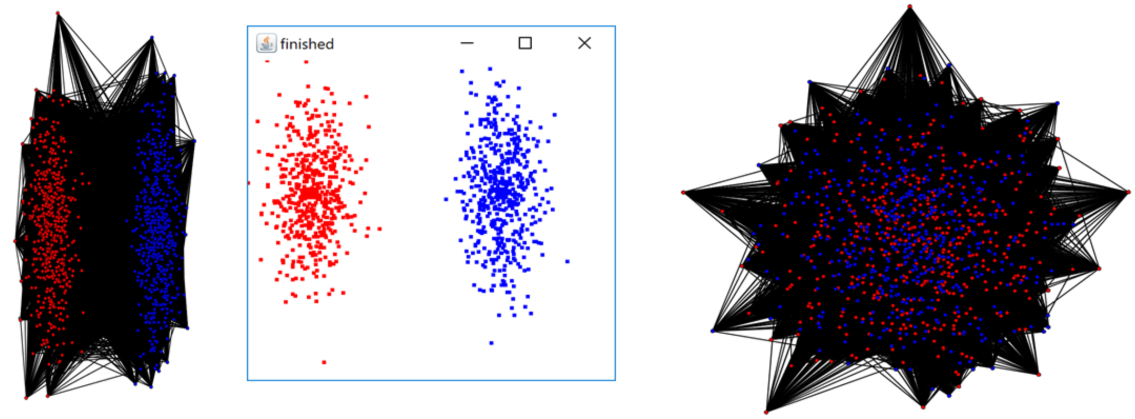
Figure 2: Interpretation of Information Content.



(a) Agreement
Of graph and categorical information
X-coordinate: all information

(b) Disagreement
Clusterd graph and random categorical attribute
X-coordinate: graph information
Y- Coordinate attribute information

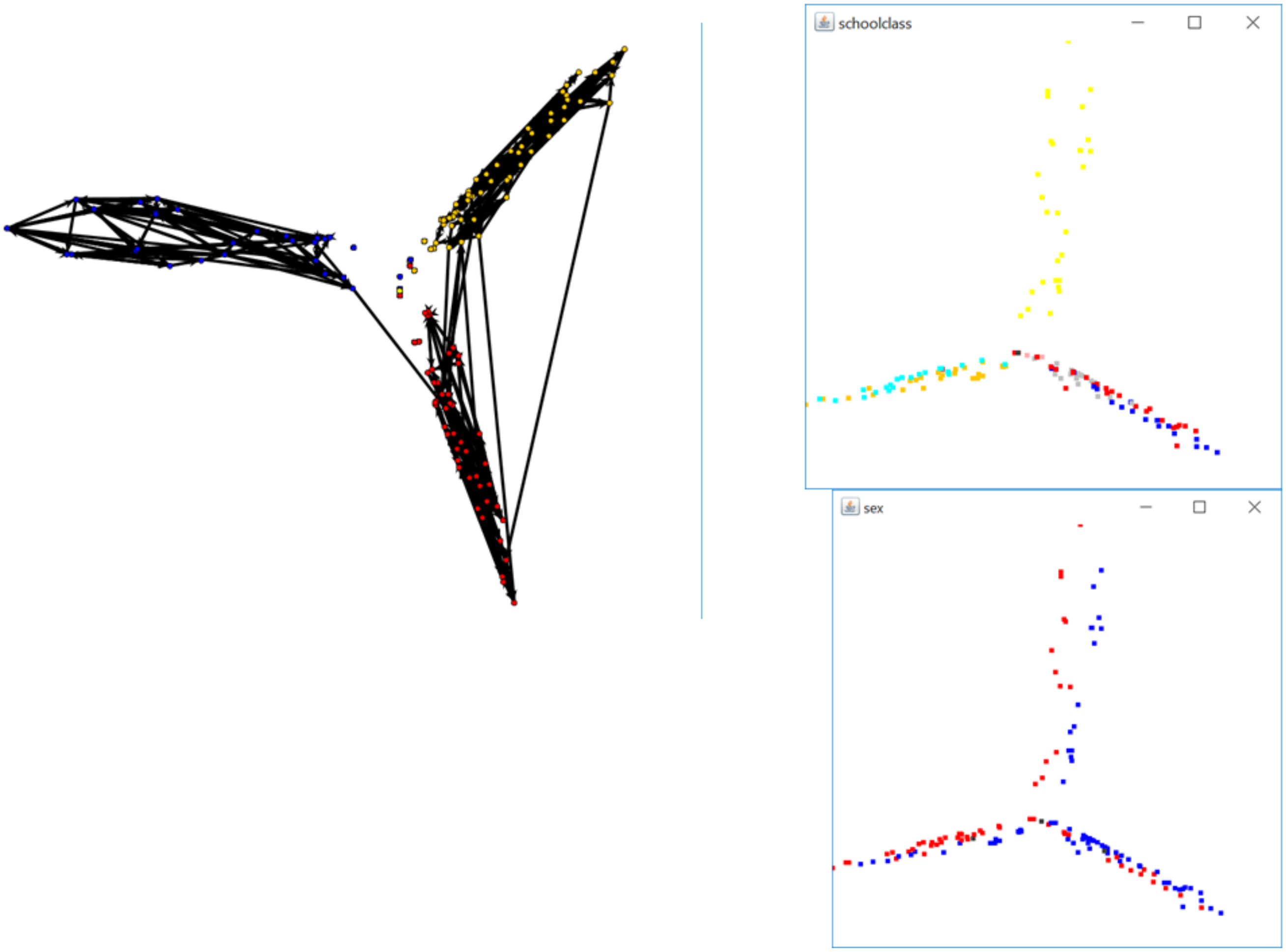
Figure 3: Interpretation of Information Content (cont.).



(c) Disagreement
Random graph and random attribute
X-Coordiante: attribute information
Y-coordinate local random structure variantions

Laplacian Eigenmap/Spectral embedding
Of the graph only; no structure visible

Figure 4: High School Friendship Network with Sex and Schoolclass.


schoolclass
sex