Unified and Contrasting Cuts in Multiple Graphs: Application to Medical Imaging Segmentation

Chia-Tung Kuo
Dept. of Computer Science
University of California, Davis
tomkuo@ucdavis.edu

Xiang Wang
Google
xiangwa@google.com

Peter Walker
Naval Medical Research Center
peter.b.walker.mil@mail.mil

Owen Carmichael
Center for Neuroscience
University of California, Davis
ocarmichael@ucdavis.edu

Jieping Ye
Dept. of Computational Medicine and Bioinformatics
Dept. of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor
jpye@umich.edy

Ian Davidson
Dept. of Computer Science
University of California, Davis
davidson@cs.ucdavis.edu

ABSTRACT

The analysis of data represented as graphs is common having wide scale applications from social networks to medical imaging. A popular analysis is to cut the graph so that the disjoint subgraphs can represent communities (for social network) or background and foreground cognitive activity (for medical imaging). An emerging setting is when *multiple* data sets (graphs) exist which opens up the opportunity for many new questions. In this paper we study two such questions: i) For a collection of graphs find a *single* cut that is good for all the graphs and ii) For two collections of graphs find a *single* cut that is good for one collection but poor for the other. We show that existing formulations of multiview, consensus and alternative clustering cannot address these questions and instead we provide novel formulations in the spectral clustering framework. We evaluate our approaches on functional magnetic resonance imaging (fMRI) data to address questions such as: "What common cognitive network does this group of individuals have?" and "What are the differences in the cognitive networks for these two groups?" We obtain useful results without the need for strong domain knowledge.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; E.1 [**Data Structures**]: Graphs and networks

© 2015 Association for Computing Machinery. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of the United States government. As such, the United States Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
KDD '15, August 10-13, 2015, Sydney, NSW, Australia.
© 2015 ACM. ISBN 978-1-4503-3664-2/15/08…$15.00.
DOI: http://dx.doi.org/10.1145/2783258.2783318

General Terms

Algorithms, Experimentation

Keywords

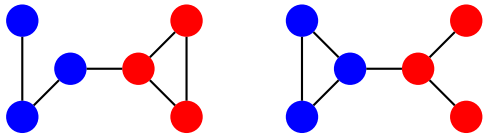
Graph cuts, fMRI, Application

1. INTRODUCTION

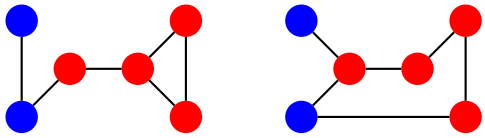
Graphs are useful in modeling the relations among a set of entities and have been extensively used in many fields. A common objective is to cut the graph to form subsets of nodes which share some topological or functional commonality, such as community detection [5], role discovery [7] or cognitive network activity [17]. A natural progression from analysis of a single graph is analysis of multiple graphs with the same node set but different topologies. This naturally exists in the medical imaging domain where in group studies we look at multiple scans of the brain at once and proper image registration ensures the scans are perfectly aligned [3].

The problem of finding cuts across *collection of graphs* poses new research questions. In this paper we examine two such questions namely finding a *single unified* cut that explains multiple graphs (a single cohort) or a *single contrast* cut which finds differences between multiple groups of graphs (different cohorts). Such scenario arises frequently in medical imaging analysis, where these images/graphs come from individuals that can be grouped into different populations by their attributes (e.g. male v.s. female, healthy v.s. demented, young v.s. elderly, etc). Figure 1 provides an visual illustration on the problem setting and the desired outputs we are looking for.

It is important to note that popular approaches such as multiview [31], consensus [18] and alternative clustering [4] that we and others have studied before are not applicable and solve different problems. Multiview clustering assumes a compatibility requirement between the views and multiview spectral methods typically combine the graphs into a single one. Consensus clustering works directly on the resultant clusterings whilst we plan to work directly on the mul-



(a) Unified



(b) Contrasting

Figure 1: An illustrative example demonstrating unified and contrasting cuts between two graphs with the same nodes. The two different colors indicate the two sides of the cut. The unified cut has a cut cost of 1 for both graphs whilst the contrast cut has a cut cost of 1 for one graph but 3 for the other.

tiple graphs. Alternative clustering finds multiple different clusterings for the one data set and hence are not applicable. Figures 7 and 8 show that consensus and multiview methods are not the most suited for our medical imaging problems.

To motivate and illustrate the difficulty of the problem consider a data set of functional magnetic resonance imaging (fMRI) data in resting state in which each scan comes from a healthy individual at rest. We may be interested in finding a unified cut from the images of healthy subjects. Applying standard spectral clustering individually to correlation graphs [28, 17] constructed from scans give widely different results as shown in Figure 2. This is due to a number of reasons but namely the underlying unreliability of the fMRI scans.

We summarize our contributions below.

- We formulate the novel problems of finding *a single unified* cut in a collection of graphs and *a single contrast* cut in two collections of graphs (sections 2.2 and 2.3).
- We show how to efficiently solve these formulations as eigendecomposition problems and discuss their interpretations (sections 2.4–2.5).
- We demonstrate the usefulness of our approach on a real world fMRI scan data set (section 4) and obtain results improving on our earlier work [3] which required strong domain knowledge.
- We demonstrate the advantages of our approach over related work such as consensus and multiview clustering for this problem (Figures 7 and 8 in section 4).

Our paper is organized as follows. In section 2 we formally introduce our problem formulation, how to solve them and discuss interpretation and parameter selection in subsections. We follow this by a discussion of related work in section 3. In section 4 we empirically evaluate our proposed approach on a real world fMRI scan data set; we demonstrate the usefulness of our approach and compare it against other recent work that might be viewed as addressing similar problems.

2. PROBLEM FORMULATION

Here we start with a brief review of spectral clustering and later subsections formally present our problem formulations, how to solve the problems and discuss their interpretations, parameter choices and possible extension.

2.1 Preliminary

We briefly review the basics of spectral clustering for completeness and to introduce notation. Readers familiar with spectral clustering can skip this subsection. We follow the standard definitions used in the literature [26].

A graph with n nodes is given as a (nonnegatively) weighted affinity matrix A of size $n \times n$ where A_{ij} is the affinity between node i and j. The degree of a node can be calculated by summing over its affinities to all the nodes and a diagonal degree matrix is defined as $D_{ii} = \sum_{j=1}^{n} A_{ij}$. The (unnormalized) graph Laplacian is defined as $L = D - A$. Spectral clustering algorithm solves a relaxed version of the min-cut problem which seeks a partition $\{C_1, \ldots, C_k\}$ of the nodes that minimizes the following.

$$\texttt{Ncut} = \frac{1}{2} \sum_{i=1}^{k} \frac{\sum_{i \in C, j \in \bar{C}} C_{ij}}{\texttt{vol}(C_i)} \tag{1}$$

where the volume is the sum of all affinities $\texttt{vol}(A) = \sum_{i,j} A_{ij}$. The relaxed problem has the following objective (for 2-way partition)

$$\begin{aligned} \underset{u}{\text{minimize}} \quad & u^T \bar{L} u \\ \text{subject to} \quad & u^T u = \texttt{vol}(A), \; u \perp D^{\frac{1}{2}} \mathbf{1} \end{aligned} \tag{2}$$

where $\bar{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$ is the normalized Laplacian. It can be shown that the optimal solution is the eigenvector corresponding to the second smallest eigenvalue of $\bar{L}$ (assuming the graph is connected). Throughout our presentation, we use L for normalized Laplacian and enforce constraint $u^T u = 1$ (instead of $u^T u = \texttt{vol}(A)$) since it merely scales all results by a fixed constant.

2.2 Unified Cut

We are given a collection of weighted graphs as affinity matrices $\{A_1, A_2, \ldots, A_k\}$ with the same node set and we wish to find a *single* unified cut that:

i Has low cut costs on all graphs and

ii The single cut is similar to each graph's individual min-cut.

The first criterion requires the output cut to *quantitatively* be good on every A_i. The enforcement of the second criterion is to make sure the output cut is *qualitatively* similar to each individual min-cut.

The cost of a cut v on a graph A is defined as in typical spectral clustering $C(v, A) = v^T L v$ where L is the normalized Laplacian of A and v is a normalized unit vector. Further we define the similarity between cuts u and v as the square of the cosine similarity $(u^T v)^2 = u^T (v v^T) u$, assuming u, v are both normalized to unit ℓ_2 norm. The reason for squaring the term is to make sure that the term is invariant to cluster ordering/relabeling (i.e. the foreground activity is cluster 1 in one clustering and 2 in another). The individual (relaxed) min-cut of each graph A_i can be computed

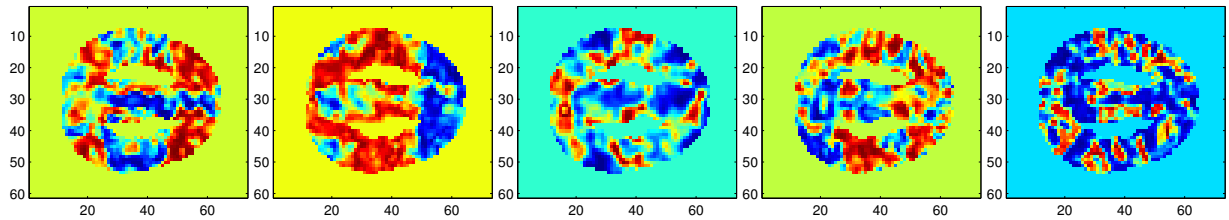


Figure 2: The individual (relaxed) normalized min-cuts for the correlation graphs constructed from fMRI scans of five healthy subjects.

as the eigenvector, v_i, of the Laplacian L_i corresponding to the second smallest eigenvalue.

Criterion (i) requires the average cut costs, assumed by the resulting cut on the graphs is small and (ii) enforces the resulting cut to be similar to each individual graph's min-cut. We note that maximizing (ii) on its own is equivalent to minimizing its negation. Thus with the two objectives (i) and (ii) above in mind, we adopt a standard approach of combining them into a single objective function with a weighting parameter α as follows.

$$\boxed{\begin{aligned} \underset{u,u^Tu=1}{\text{minimize}} \quad & \frac{1}{k}\sum_{i=1}^{k} u^T L_i u - \alpha(\sum_{i=1}^{k}(u^T v_i)^2) \\ & = u^T \left[\sum_{i=1}^{k}(\frac{1}{k}L_i - \alpha v_i v_i^T)\right] u \end{aligned}} \tag{3}$$

2.3 Contrast Cut

In the contrast cut setting we are given two collections of weighted graphs with the same node set (e.g. fMRI scans taken from the same imaging device). We aim to find a single discriminative cut between the two collections in that the cut is small for the first collection but large for the second. As before we look for a (relaxed) cut but here seek a cut whose costs on the individual graphs best distinguish the two collections. Note that in this case we do not enforce the contrast cut to be necessarily similar to any individual min-cuts.

Our objective function is defined as follows:

$$\boxed{\underset{w,w^Tw=1}{\text{minimize}} \quad w^T(\frac{1}{k}\sum_{i=1}^{k} L_i^{(A)} - \beta\frac{1}{m}\sum_{j=1}^{m} L_j^{(B)})w} \tag{4}$$

where the two collections of (edge-weighted) graphs with the same node set are given as affinity matrices $\{A_1, A_2, \ldots, A_k\}$ and $\{B_1, B_2, \ldots, B_m\}$. $L_i^{(A)}$ and $L_j^{(B)}$ denote the graph Laplacians of A_i and B_j, respectively, and β is a weighting parameter.

The idea of using combinations of multiple Laplacians has been explored before but in different problems. [24, 1] both use a convex combination of Laplacians along with another loss function in a semi-supervised setting. In our case since we assume these Laplacians are constructed from entities of the same modality and thus compatible within each class, we simply weigh them equally resulting in a linear combination. If other source of knowledge is available, other regularization terms can be added to calibrate the weighting among the given Laplacians.

2.4 Interpretation of Two Objective Functions

In this section we delve deeper into the objective functions and discuss their interpretations.

2.4.1 *Maximizing Squared Cosine Similarity*

In the second part of the unified objective function we maximize the squared cosine distance. Each relaxed cut v_i for L_i can be viewed as a vector in the space $\mathbb{R}^n$. A natural similarity measure in such a space is the cosine similarity, which characterizes the size of the angle between two vectors. In our case where the vectors are all normalized, maximizing the cosine similarity to a given vector is equivalent to minimizing its Euclidean distance as shown below.

$$\begin{aligned} & \underset{u,u^Tu=1}{\text{minimize}} \quad ||u-v||_2^2 = u^Tu - 2u^Tv + v^Tv \\ \equiv & \underset{u,u^Tu=1}{\text{minimize}} \quad -u^Tv \\ \equiv & \underset{u,u^Tu=1}{\text{maximize}} \quad u^Tv = \frac{u^Tv}{||u||_2||v||_2} \end{aligned} \tag{5}$$

where the second and third lines each follow because u and v have unit ℓ_2 norm and thus u^Tu and v^Tv are both constant in the objective. Since the vectors are arbitrary up to a multiple of ± 1 without affecting the objective value, squaring the cosine similarity makes sign flipping irrelevant and therefore avoids the need to pre-align those cut vectors.

2.4.2 *Summation and Subtraction of Laplacians*

Here we give an interpretation of the summation and subtraction involving Laplacians in our objective functions in terms of the graph affinities. In equation 3, we have a sum of terms of the form $\frac{1}{k}L_i - \alpha v_i v_i^T$ where v_i is the relaxed optimal min-cut for L_i. If we treat v_i as the ground truth partition for A_i then the outer-product $v_i v_i^T$ can be viewed as the ideal similarity matrix that would result in such partition and A_i as a perturbation of it. By the definition of the normalized Laplacian, $L_i = I - D_i^{-\frac{1}{2}} A_i D_i^{-\frac{1}{2}}$, it is the identity matrix minus the degree normalized affinity matrix. Accordingly $\frac{1}{k}L_i - \alpha v_i v_i^T = \frac{1}{k}\left[I - (D_i^{-\frac{1}{2}} A_i D_i^{-\frac{1}{2}} + \alpha k v_i v_i^T)\right]$ can be viewed as complementing the original affinity A_i with the knowledge from an idealized similarity $v_i v_i^T$ that would result in the current best partition. We are therefore using the individuals' min-cuts as "ground truth" and then perturb each graph towards its ground truth. Notice however that since $v_i v_i^T$ contains both positive (i.e. similar) and negative (i.e. dissimilar) entries, the perturbed similarity will typically have negative entries and $\frac{1}{k}L_i - \alpha v_i v_i^T$ need not be positive semidefinite. The effect of this is that a slightly more

complex method of finding the appropriate eigen-vector is needed as discussed in section 2.5.

For both objectives in equations 3 and 4, averaging Laplacians is involved. Averaging multiple Laplacian is similar to averaging the original affinity matrices and then creating a single Laplacian from it if the nodes in different graphs have similar degrees. In fact when the degree matrix is identical for different graphs, averaging Laplacians is exactly equivalent to averaging affinity matrices. In addition, the objective of contrast cut (4) also involves the subtraction of Laplacians. The effect of subtracting a Laplacian from another, as in the contrast case, can be viewed in a similar fashion. If two graphs have approximately the same degrees ($D_i \approx D_j \approx D$), then

$$\begin{aligned} L_i - \beta L_j &= I - D_i^{-\frac{1}{2}} A_i D_i^{-\frac{1}{2}} - \beta(I - D_j^{-\frac{1}{2}} A_j D_j^{-\frac{1}{2}}) \\ &\approx (1-\beta)[I - D^{-\frac{1}{2}}(\frac{A_i - \beta A_j}{1-\beta})D^{-\frac{1}{2}}] \end{aligned} \tag{6}$$

The result is as if we define a new similarity where the affinity defined by A_j counteracts affinity A_i. That is, a pair of nodes considered similar in A_j drives the pair more dissimilar in the resulting affinity, and vice versa. Again due to the subtraction and its resulting negative entries, $L_i - \beta L_j$ will not be positive semidefinite and the eigenvectors are chosen as shown in section 2.5.

2.5 Solution and Related Issues

In this section we solve the proposed optimization problems and discuss some related issues with the use of the solutions and their extensions.

Solving eigenvalue problem. Since the objectives have the same form (i.e. $u^T M u$ for some constant M), we describe our solution to only the unified cut formulation and similar derivation follows for the contrast case as well. The Karush-Kuhn-Tucker (KKT) conditions [21] are commonly used in solving constrained optimization problems with differentiable objective functions. We follow the standard technique and define the Lagrangian as

$$\mathcal{L} = u^T M u - \lambda(u^T u - 1) \tag{7}$$

where $M = \sum_{i=1}^{k}(\frac{1}{k}L_i - \alpha v_i v_i^T)$ is a constant (in the contrast cut case, $M = \frac{1}{k}\sum_{i=1}^{k} L_i^{(A)} - \beta\frac{1}{m}\sum_{j=1}^{m} L_j^{(B)}$). The first order KKT conditions then require the following to hold on any solution u^*.

$$\begin{aligned} &Mu^* - \lambda u^* = 0 && \text{(stationarity)} \\ &{u^*}^T u^* = 1 && \text{(primal feasibility)} \end{aligned} \tag{8}$$

Notice that we have only two conditions above since the dual feasibility and complementarity conditions are only applicable when inequality constraints are present. From here we can clearly see that the candidate feasible solutions are the eigenvectors of M and the eigenvector corresponding to the smallest (numeric) eigenvalue is the desired solution. Notice that in each formulation, the matrix M is symmetric . This guarantees that all eigenvalues are real-valued and thus finding the smallest numeric eigenvalue is unambiguous. Such eigenvalue problem can be readily solved for by algorithms implemented in modern software (e.g. `eigs` in MATLAB).

Avoiding close-to-trivial solution. An additional step is needed after we compute the smallest eigen-pairs. In the standard spectral clustering formulation, the eigenvector corresponding to the eigenvalue 0 will represent a trivial cut (i.e. having all nodes on one side and none on the other). This is dealt with by choosing from the second and later eigenvectors. In terms of the optimization problem, an orthogonality constraint is enforced.

$$u^T D^{\frac{1}{2}} \mathbf{1} = 0 \tag{9}$$

where $\mathbf{1}$ is the vector having 1 in every entry. However the subtractions in the objectives (and hence the resulting M) make the first few eigenvalues become negative and there will unlikely be any eigenvectors satisfying the constraint above. Instead we still want to eliminate eigenvectors that are sufficiently close to the trivial solution and thus select only from those satisfying the following

$$||u^T D^{\frac{1}{2}} \mathbf{1}|| \leq \epsilon ||D^{\frac{1}{2}} \mathbf{1}|| \tag{10}$$

where D is the degree matrix of $\frac{1}{k}\sum_{i=1}^{k} A_i$ and ϵ is a tolerance between 0 and 1. This non-triviality check is only needed when the weighting parameter α or β is close to 0. When they are set to moderate sizes, none of the eigenvectors would resemble the trivial solution in the standard spectral clustering. We summarize the solution steps concisely in Algorithm 1.

Algorithm 1: Pseudocode for the solution steps

Input: $\{L_1, \ldots, L_k\}$ (and $\{L_1^{(B)}, \ldots, L_m^{(B)}\}$ in the contrast case)
Output: u

```
M ← Σ_{i=1}^{k} (1/k L_i − α v_i v_i^T) (or in the contrast case,
  M ← 1/k Σ_{i=1}^{k} L_i − β 1/m Σ_{j=1}^{m} L_j^(B));
Compute the q eigenvectors {u_1, ..., u_q} corresponding
  to the smallest q eigenvalues to Mu = λu;
for i ← 1 to q do
    if ||u_i^T D^{1/2} 1|| ≤ ε||D^{1/2} 1|| then
        u ← u_i;
        break;
    end
end
return u;
```

Extension to multi-way cuts. As in the standard spectral clustering, a multi-way cut (i.e. partition) can be obtained by examining multiple eigenvectors. A (hard) partition of the nodes can then be decided based on techniques such as the eigengap heuristic [26] or the closeness of the eigenspace to a canonical coordinate system [30]. In a similar manner, we could also utilize multiple eigenvectors in our formulations: for each graph Laplacian we compute its top q eigenvectors and stack them columnwise into V_i in place of v_i in formulation (3). In this sense the unity term (i.e. second term in equation (3)) encourages small angular distances between each pair of bases of the eigenspaces. One point worth mentioning is the possibility of repeated eigenvalues (or very close eigenvalues). In such cases, any rotation of the eigenvectors (i.e. an orthogonal transformation) gives the same eigenspace. This however does not cause ambiguity in our formulation since we only utilize the outer products of the eigenvectors and any rotation of the coordinates gives the same result. Specifically if $\hat{V} = VR$ where R is a $q \times q$ orthogonal matrix, then $\hat{V}\hat{V}^T = VRR^TV^T = VV^T$.

2.6 Choosing Weighting Parameters

Here we discuss the selection of the weighting parameters α and β used in our experiments. The parameters α and β are used to control the trade-off between two objective terms as in many optimizations with regularization, such as sparsity-inducing, smoothing, etc. Unlike classification or regression problems where one can set parameters using cross-fold validation there does not seem to exist a well accepted principled approach to such selection in unsupervised learning. Instead we perform an analogous procedure by optimizing the desirable property of each resulting cluster in a partition being dense and well separated from others. Specifically in our experiments, we run kmeans on our solution eigenvectors for each setting of α (or β), and choose the value that gives the smallest sum of within-cluster distances. We note there exist other approaches in evaluating the clustering quality such as the information theoretic approach based on minimum description length principle [25, 11].

3. RELATED WORK

Several bodies of work are related to our proposed research that fall into the category of analyzing multiple graphs.

Multiview spectral clustering has been well studied by recent work [15, 14, 31]. As mentioned in the introduction this line of work may appear to be similar to our unified problem setting but not contrasting case. One such formulation [15] outputs a cut for each graph instead of a single cut as our work does. The same set of authors [14] proposed an iterative algorithm that utilizes eigenvectors from different Laplacian views to modify each graph. Both algorithms [15, 14] iterate on pairs of graphs during the iterative optimization steps and thus can be inefficient when the number of views is much more than two, as is the case in our targeted application. Perhaps the closest work to our own is [31] which combines multiple views into one and performs a spectral cut on it. We empirically investigate this method (see Figure 8) and discuss it in more detail in section 4.

Consensus clustering investigates the problem where a data set can be partitioned into multiple different clusterings and one wishes to find a single clustering which is most similar to all the given clusterings. These different clusterings can arise because, for instance, different clustering algorithms are applied to the data or different initializations are used for the same algorithm [25, 18]. The problem is formally defined as a median partition problem which minimizes (maximizes) the sum of distances (similarities) between the single consensus clustering and the rest of all given clusterings [8]. For most commonly used distance/similarity between clusterings (e.g. symmetric difference or normalized mutual information) this optimization problem is NP-complete [27] and greedy and local search heuristics are used to find quality solutions [18].

There are core differences between our work and consensus clustering. Consensus clustering was motivated by finding agreement between multiple different clusterings for the one data set whereas we are focusing on finding a unified clustering from multiple graphs/datasets. Furthermore, as alluded to in the introduction, consensus clustering is a distinct two-step process where the first step finds multiple quality partitions using different methods. This second step is formulated without regard to the first step, i.e. the generation of multiple partitions. Our unified cut formulation, on the other hand, *simulataneously* looks for a high quality and common partition. We demonstrate the advantage of our formulation in our experiment against consensus clustering (see Figure 5 v.s. Figure 7).

Alternative clustering aims to find alternative partitions different from some given partition but with approximately the same objective function value as the first clustering [4, 22]. Some work [20] defines new objectives that incorporates terms on the novelty relative to previous discovered partitions. Other work [13, 12] attempts to find disparate clusterings simultaneously; that is finding multiple distinct clusterings on the same data. Our problem setting finds a single contrast cut on multiple collections of graphs unlike alternative clustering that finds multiple clusterings on just one data set.

Graph mining looks for frequent substructures (e.g. connected subgraphs) in a database of graphs [29]. However these are not applicable to the medical imaging data which introduces additional challenges that limit the practicality of these methods. For example, these methods typically only work well when the graph database is large enough whereas when analyzing medical imaging data we often have a small number of graphs (typically 10s or at most 100s). In addition our work is with graphs with weighted edges, unlike graph mining methods which typically assume edges are unweighted.

Contrast mining in general studies the problem of finding discriminative patterns between classes of data, and it encompasses many fields in data mining such as classification, times series analysis (e.g. change point detection), and market-basket analysis (e.g. class-based association rules), etc [2]. In the context of contrast graph mining [23], the goal is to find patterns (i.e. subgraphs) frequent in one database and infrequent in another. The input and output are similar to the case of graph mining above. Our work is in a different context as described above.

Our earlier work on medical imaging has studied problems such as finding cognitive networks in fMRI scans (e.g. shapes of the regions and temporal activation) with constrained tensor factorization [3, 16] and finding discriminative patterns between subjects belonging to different groups by adjusting the nodes in the graphs before clustering [17]. However, unlike our current work, this prior work either studies single scans with significant guidance [3, 16], or when studying multiple scans, looks for multiple individual best cuts instead a single one [17].

4. EMPIRICAL EVALUATION

Here we empirically evaluate our approaches on a real world data set of fMRI scans. Specifically we aim to explore each of the following questions:

1. Does our approach find a meaningful unified cut when it is expected to exist? (e.g. among a group of normal healthy subjects)

2. Does our approach find any meaningful unified cut when it is *not* expected? (e.g. among a group of demented subjects)

3. Does there exist a contrast cut to distinguish the two distinct groups? (e.g. contrast between the normal and the demented subjects)

4. How does our unified cut approach compare to existing multiview and concensus clustering approaches?

In the following subsections, we describe the data set used and the preprocessing step. Then we discuss the results of our approach as well as some other existing work on one major task in medical imaging analysis.[1]

4.1 Data Description and Preprocessing

In these experiments we evaluate our approaches on a data set of resting state functional magnetic resonance imaging (fMRI) scans. This data set is available from the U.C. Davis Alzheimer's institute after signing the appropriate privacy disclosures since the data is of real patients. Our data set consists of scans from 61 subjects, of which 19 were normal subjects, 21 were mildly cognitively impaired (MCI), and the remaining 21 were diagnosed as demented. Each scan is a series of snapshots of 3D brain images of size $61 \times 73 \times 61$ over ~ 200 time steps. For ease of presentation in 2D we use a middle slice of these scans. In addition because these images were pre-aligned (registered), we use a brain-shaped mask (provided by neurology professionals) to work with only the voxels that are considered parts of the brain. Overall each of the scans is represented by ~ 1700 voxels. For each scan we construct its affinity graph where the affinity between node i and node j is the absolute-valued Pearson's correlation coefficient between the activation time series of the i-th and j-th voxels. Such measurements have been widely used with success in neuroimaging community [6] and in our earlier work [17].

4.2 Major Tasks

fMRI scans can be viewed as exhibiting a complex interactions of neural signals and background noise. Graph cuts (or more generally image segmentation techniques) can be viewed as trying to separate the foreground activity from the background noise. A major task among the studies of neuroimaging is to identify the structural differences between a healthy individual and that of a demented individual and it is natural to examine if graph cuts can successfully identify any discriminative patterns. A well known cognitive network believed to be present in all healthy people when completely at rest is the default mode network (DMN). Neurological studies had suggested that the strength of DMN can be a strong indicator for the diagnosis of Alzheimer's disease [9]. An idealized DMN is shown in Figure 3. As shown in Figure 2, due to the noise in fMRI scans, min-cuts on each individual healthy scan can vary greatly and none of them resembles the idealized DMN. In our earlier work [3, 17] we were able to discover the DMN *by using strong prior guidance*. Here no such guidance is used.

We perform the following experiments with our formulations to address the first three questions.

1. Find a unified cut in the normal group.
2. Find a unified cut in the demented group.
3. Find a contrast cut between the normal and demented groups.

For each experiment, we test a range of values for the tradeoff parameters α and β and select them using our approach

[1]Our source code is publicly available at `https://sites.google.com/site/chiatungkuo/`.

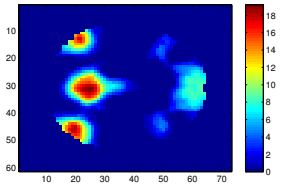


Figure 3: Idealized default mode network (DMN).

described in section 2.6. **It is worth noting that we do not use the MCI subjects to determine the cut but also evaluate the cut on them in the following section.** MCI subjects exhibit onset of symptoms believed to be linked to dementia. For the last experiment question we compare our results with one existing multiview spectral clustering [14] and another consensus clustering approach [18] designed for graphs.

4.3 Measurements

Here we outline the results we present for each experment. For each experiment, we run our unified or contrast clustering algorithm to find a single cut. The measurements we provide are typical in the medical imaging literature and consists of three results for each experiment. First, we can score the quality of the single cut against a variety of instances within the different sub-populations and represent the results as a box plot with a box for each of the three sub-populations of subjects (i.e. normal, MCI, and demented). As mentioned earlier it is well known that normal subjects should exhibit a common underlying cut whilst other sub-populations should not. Secondly, we use an independent two-sample T-test on the means of the cut costs for each pairing of the three sub-populations. The T-test scores provide a significance test on the null hypothesis that the grouped samples actually come from the same distribution; therefore ideally we would like to see this hypothesis rejected (e.g. p-value < 0.05). Finally, we present the actual cut as a color coded image with blue and red being on different sides of the cut and far from the cut. We have colored these cuts so that the foreground activity is in red.

4.4 Results and Discussion

Figures 5, 4 and 6 show the results for each of the three settings above, respectively.

Question 1: Unified Cut of the Normal Group. Here we attempt to see if our method finds a reasonable unified cut for all normal subjects. As can be seen from Figure 5, the unified cut from the group of normal subjects has close visual resemblance to the default mode network (Figure 3). This is in strong agreement with the phenomenon observed in neuroscience community [9] which postulates the existence of such cognitive network in all healthy subjects. Furthermore Figures 5(b) and 5(c) demonstrate that learning a unified cut on the normal subjects **alone without knowledge of MCI or demented subjects** is able to identify statistically significant (i.e. T-test) discriminative cut. In contrast our earlier work [3] was only able to identify the DMN with significant guidance. The lack of discrimination between the MCI and demented subjects, however, is not too surprising; not only do we learn a unity on the normal

subjects alone, but the MCI subjects themselves are actually all believed to eventually deteriorate to be demented.

Question 2: Unified Cut of the Demented Group. The results for a unified cut of the demented people is shown in Figure 4. The unified cut found is similar to a bisection of the brain in the middle and exhibits no clear discernible pattern. This result is in fact expected as dementia could be consequences of deformations or impairment of many possible brain regions; no compelling reason is known to enforce a group of demented people to share similarly discernible networks on their brain scans apart from non-demented counterparts. Furthermore as expected, compared to the unified results from the normal sub-population, the grouped cut costs induced by the unified cut of the demented cannot differentiate sub-populations significantly (Figures 4(b) and 4(c)) as far as can be demonstrated by the T-test (p value > 0.5).

Question 3: Contrast between the normal and demented. We show the results of our contrast cut method between the normal group and the demented group in Figure 6. The foreground in the contrast cut (Figure 6(a)) contains part of the DMN and can be interpreted as the parts of the DMN that funcion *together in normal patients but not in the demented patients*. One explanation is that the demented subjects still exhibit a partial DMN as it is still intact but deteriorating [6, 9]. Here our objective function explicitly demands low cut costs on the normal and high cut costs on the demented and therefore only a portion of DMN is discovered from the cut. To test the objective, the boxplot (Figure 6(b)) and the corresponding T-tests (Figure 6(c)) demonstrate that the normal group is well distinguished from the MCI and the demented. Importantly, our method can now differentiate the MCI and the demented groups at the 95% confidence result. **This was not possible with our earlier work even with strong guidance [3, 17] since the guidance was in the form of what to expect in normal subjects.**

In the following subsections we address question 4 via comparing the results from our unified setting (Figure 5) to two other methods: one adapted from a recent consensus graph clustering method [18] and the other is a multiview spectral clustering method [31]. The contrast cut setting is novel and no standard comparisons are apparent to our knowledge.

4.5 Question 4: Comparison to Consensus and Multiview Clustering Approaches [18] [31]

Consensus Clustering. Here we redo our experiments on unified cuts on the normal group but instead use the competing method of finding a consensus partition from clustering each individual scan separately. Note this is a different problem setting than our own since the input into consensus clustering algorithm is just the individual clusterings from each scan but not the Laplacians. However, the clusterings given to the consensus algorithm **are the same** as that used our in own method, i.e. those cuts used in the second expression of equation 3.

The motivation for this method is for a single data set, multiple partitions were either returned from different clustering algorithms or from multiple runs of one clustering algorithms with different initializations. There are many consensus clustering algorithms and we chose one whose results are most promising and most applicable to our problem setting. This method was published in *Science* and is directly applicable to finding a consensus network/graph. The general idea behind this method [18] is to construct a new consensus matrix based on the multiple partitions. The consensus matrix can essentially be viewed as a new similarity matrix where the similarity between a pair of nodes is the proportion of times they fall into the same cluster among the given partitions. Afterwards the same clustering algorithm(s) is run on this consensus matrix to generate yet another set of multiple partitions. This process was iterated until the consensus from one iteration is the same as in the previous iteration (or close enough within some pre-determined tolerance).

In our context we can view the individual min-cuts from each scan as the multiple partitions of the same set of nodes. We construct a consensus matrix as described above where two nodes are considered to fall in the same cluster if they have the same sign in the eigenvector (i.e. viewed as a 2-way partition). We use this consensus matrix as the new affinity between nodes and compute the min-cut from this affinity matrix with standard spectral clustering. Since spectral clustering always returns the same min-cut, there is no need to iterate the process and we simply output this min-cut on the consensus affinity matrix as the consensus partition. Our adapted version is presented in pseudocode in Algorithm 2.

Figure 7 shows the results from this adapted approach. This approach does differentiate cut costs between normal and demented subjects with reasonably small p-values, though not as extreme as in our unified cut approach. We do however point out that low cut cost is the objective of good clusterings and since there is no *enforcement of collective low costs*, the cut costs are in general higher than those in our approach (see left boxe in Figures 5(b) and 7(b)). Similarly the cut shares, to certain extent, visual closeness to Figure 5(a) and the DMN, but the regions are not as distinctly defined as in Figure 5(a).

Algorithm 2: A consensus clustering algorithm adapted from the approach described in [18]

Input: k individual min-cuts $\{v_1, \ldots, v_k\}$
Output: a single consensus clustering u

```
Initialize C ← n × n matrix of zeros, where n is the
  number of nodes;
for i ← 1 to n do
    for j ← 1 to n do
        for m ← 1 to k do
            if v_m(i)v_m(j) ≥ 0 then
                C(i, j) ← C(i, j) + 1;
            end
        end
    end
end
A ← C/n;
Compute the min-cut u of graph affinity A using
  regular spectral clustering;
```

Multiview Spectral Clustering. The work of [31] is one of the most popular multiview spectral clustering approach and, as noted in the Related Work section, its output is most directly comparable to our unified cut setting. This work addresses the setting of multiple *directed* graphs

and formulates the problem from the view point of a random walk. They construct a mixture transition matrix from all underlying graphs by noting that the mixture of random walks is also another random walk (i.e. a linear combination ergodic Markov chains is another Markov chain). This is a well known result obtained by combining individual transition probabilities defined by each graph, weighted by their respective stationary distribution (and a user-supplied additional weight to each graph). Note that the following equivalence between the normalized cut and the random walk had been established [26, 19]

$$\texttt{Ncut}(A, \bar{A}) = p(A \to \bar{A}|A) + p(\bar{A} \to A|\bar{A}) \qquad (11)$$

It provides a nice interpretation to minimizing the normalized cut: it is minimizing the probability that a random walker in the next step would transition across the two subsets, either from A to $\bar{A}$ or from $\bar{A}$ to A. The work in [31] has a analogous interpretation as random walk where in each step the walker, with certain probabilities, either continues on the current graph or jumps to one of the other graphs.

Since we do not have prior knowledge about which patient's scan is more important, we supply equal weights to each graph (as in our unified cut approach we simply average the individual relaxed min-cut objectives). The results from the multivew method [31] on the normal subjects are shown in Figure 8 (again as relaxed partition as in our other experiments). We can see that the results from this method do not produce very meaningful insights either in terms of the cut itself or the grouped boxplot.

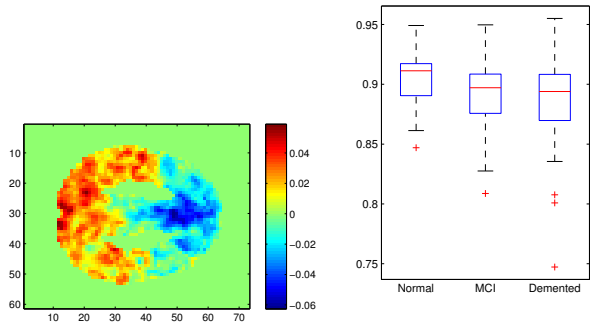
One reason behind this poor result is that the multiview setting makes the fundamental assumption of compatibility. In our context it means a good cut form one graph will also be a reasonably good cut for another, but clearly from Figure 1 this is not the case. The wide variability in the volume of our absolute correlation graphs constructed from fMRI scans is also evidence of this. In fact the volumes are directly translated to the weightings among graphs in this work [31]. Table 1 provides some statistics of the graphs and individual min-cuts from the normal subjects.

5. CONCLUSION AND FUTURE WORK

Though there has been much work on analysis of graphs, most work focuses on a single graph. In this paper we explore the area of finding a *single* cut on a collection of graphs which is a good cut for all graphs and also similar to the individual cuts for underlying graphs in the collection. We call this problem finding a unified clustering. Though this appears to be similar to the consensus and multiview clustering setting, we demonstrate the differences by our comparison against those methods (see Figures 5, 7 and 8). This is because the underlying inputs into the problem are different: whereas our method takes as input both the Laplacians and the underlying individual cuts, multiview clustering takes only the former and consensus clustering only the latter.

In this paper we also proposed the novel problem of finding a *single* contrast cut. Here a contrast cut is one that is a good cut for one collection of graphs but a poor cut for another collection of graphs. To our knowledge this problem is novel and has not been studied by the theory or applied community.

The idea of finding cuts on medical imaging is a popular method of separating out network activity (one side of the cut) to background activity (the other side of the cut). We



(a) Unified cut (b) Grouped cut costs

	MCI		Demented	
	t statistic	p value	t statistic	p value
Normal	1.3966	0.1706	1.6539	0.1064
MCI			0.5835	0.5629

(c) T test between groups (* next to p values indicates < 0.05)

Figure 4: Unity within the demented subjects

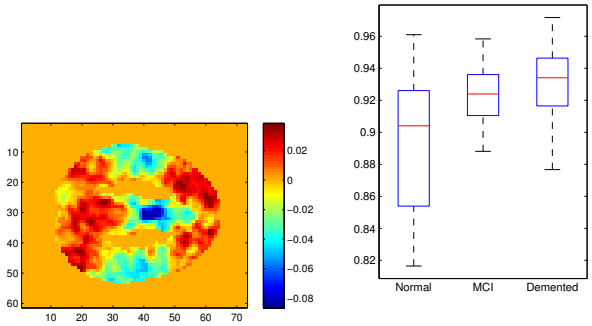
have already explored this application in earlier work [17, 28] but only for a single graph. However, in many studies there are multiple cohorts of individuals with other shared properties (i.e. normal or demented). A natural question to ask is then: "For a cohort of normal patients, what common network exists for them?" We can address this question using unified cuts. Another question to ask is, "For a cohort of normal patients and another of demented patients, what differentiates their networks?" We can address this question using contrast cuts.

We tried our method on the fMRI data set from UC Davis Alzheimer's institute (available with appropriated signed disclosures). We have used this data set in our earlier studies [3, 28, 16] but all this previous work was on analysis of a single scan. Compared to this prior work, we were able to discover the default mode network in normal patients *without* the need for strong guidance in the form of constraints. Furthermore, we were able to successfully differentiate MCI patient from demented patients, the significance of which is that MCI patients eventually develop dementia but were never used to guide the cut selection. We also addressed the novel question of how the networks in normal and demented groups differ.

Although we focus on a particular application in medical imaging in this paper, we expect our approaches to be applicable in the analysis of other data sets as well. For example, our setting is suited to the analysis of a heterogeneous network where multiple edge types exist among a single set of nodes. These networks are common in the studies of social networks where the different types of edges characterize different relationships among individuals (e.g. friendship, collaboration, communication, etc) [10]. These different edges could also possibly arise from different metrics used to measure the similarity between pairs of nodes. In either case we can form multiple graphs where edges in one graph are of a single type and apply our unified and/or contrasting cuts to learn collectively from these graphs.

	Mean	Std. dev.	Minimum	Maximum
Volume ($\times 10^6$)	0.9167	0.4312	0.06254	1.7516
Min-cut cost	0.7001	0.0416	0.6340	0.7937

Table 1: Statistics of the graphs and min-cuts from the normal subjects. This explains the poor performance of using multiview spectral clustering [31] and also demonstrates it's appropriate for us to combine objectives in equation 3.

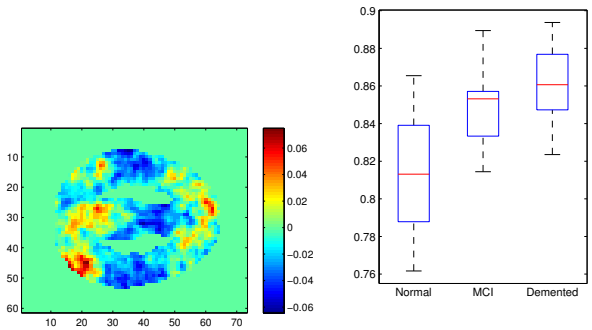


(a) Unified cut (b) Grouped cut costs

	MCI		Demented	
	t statistic	p value	t statistic	p value
Normal	-3.0384	0.0043*	-3.3349	0.0019*
MCI			-0.7930	0.4325

(c) T test between groups (* next to p values indicates < 0.05)

Figure 5: Unity within the normal subjects using our formulation. To be compared with Figure 7 and Figure 8.

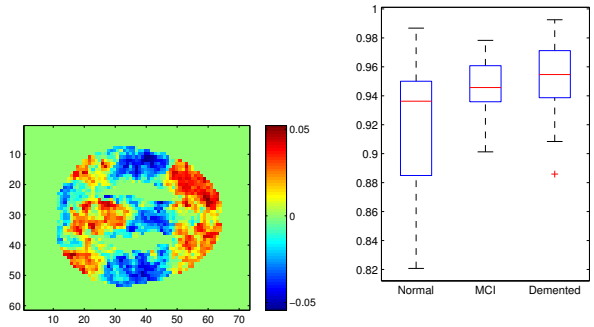


(a) Contrast cut (b) Grouped cut costs

	MCI		Demented	
	t statistic	p value	t statistic	p value
Normal	-4.0681	0.00023*	-5.7794	0.0000011*
MCI			-2.2049	0.0333*

(c) T test between groups (* next to p values indicates < 0.05)

Figure 6: Contrast between the normal and the demented subjects. The boxplot shows much stronger discriminative result than using unified cut on the normal subjects alone.

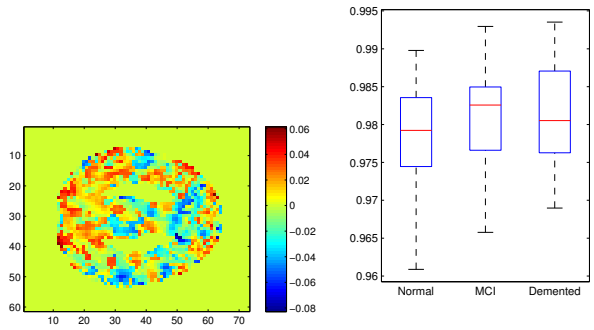


(a) Unified cut (b) Grouped cut costs

	MCI		Demented	
	t statistic	p value	t statistic	p value
Normal	-2.1484	0.0381*	-2.6533	0.0116*
MCI			-1.0258	0.3112

(c) T test between groups (* next to p values indicates < 0.05)

Figure 7: Unity within the normal subjects using an adapted consensus clustering approach from [18] outlined in Algorithm 2. To be compared against Figure 5.



(a) Unified cut (b) Grouped cut costs

	MCI		Demented	
	t statistic	p value	t statistic	p value
Elderly	-1.1042	0.2764	-1.2010	0.2372
MCI			-0.1896	0.8506

(c) T test between groups (* next to p values indicates < 0.05)

Figure 8: Multiview spectral clustering [31] on normal subjects. To be compared against Figure 5.

Acknowledgment

This research is supported by NSF Grant IIS-1422218 and ONR grant NAVY 00014-09-1-0712. The opinions of the authors do not necessarily reflect those of the United States Navy or the University of California, Davis.

Peter B. Walker is a military service member. This work was prepared as part of his official duties. Title 17 U.S.C. 101 defines U.S. Government work as a work prepared by a military service member or employee of the U.S. Government as part of that person's official duties.

The data used in the experiments were collected using NIH grants P30 AG010129 and K01 AG 030514.

6. REFERENCES

[1] A. Argyriou, M. Herbster, and M. Pontil. Combining graph laplacians for semi–supervised learning. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *NIPS*, pages 67–74. MIT Press, 2006.

[2] J. Bailey and G. Dong. Contrast data mining: Methods and applications. *IEEE ICDM 2007 Tutorials*, 2007.

[3] I. Davidson, S. Gilpin, O. Carmichael, and P. Walker. Network discovery via constrained tensor analysis of fmri data. In *ACM SIGKDD*, pages 194–202. ACM, 2013.

[4] I. Davidson and Z. Qi. Finding alternative clusterings using constraints. In *ICDM*, pages 773–778. IEEE, 2008.

[5] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[6] K. J. Friston. Functional and effective connectivity: a review. *Brain connectivity*, 1(1):13–36, 2011.

[7] S. Gilpin, T. Eliassi-Rad, and I. Davidson. Guided learning for role discovery (glrd): framework, algorithms, and applications. In *ACM SIGKDD*, pages 113–121. ACM, 2013.

[8] A. Goder and V. Filkov. Consensus clustering algorithms: Comparison and refinement. In *ALENEX*, volume 8, pages 109–117. SIAM, 2008.

[9] M. D. Greicius, G. Srivastava, A. L. Reiss, and V. Menon. Default-mode network activity distinguishes alzheimer's disease from healthy aging: evidence from functional mri. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13):4637–4642, 2004.

[10] J. Han. Mining heterogeneous information networks by exploring the power of links. In *Discovery Science*, pages 13–30. Springer, 2009.

[11] M. H. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.

[12] M. S. Hossain, S. Tadepalli, L. T. Watson, I. Davidson, R. F. Helm, and N. Ramakrishnan. Unifying dependent clustering and disparate clustering for non-homogeneous data. In *ACM SIGKDD*, pages 593–602. ACM, 2010.

[13] P. Jain, R. Meka, and I. S. Dhillon. Simultaneous unsupervised learning of disparate clusterings. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 1(3):195–210, 2008.

[14] A. Kumar and H. Daumé. A co-training approach for multi-view spectral clustering. In *ICML*, pages 393–400, 2011.

[15] A. Kumar, P. Rai, and H. Daume. Co-regularized multi-view spectral clustering. In *NIPS*, pages 1413–1421, 2011.

[16] C.-T. Kuo and I. Davidson. Directed interpretable discovery in tensors with sparse projection. In *SIAM Data Mining*, pages 848–856, 2014.

[17] C.-T. Kuo, P. B. Walker, O. Carmichael, and I. Davidson. Spectral clustering for medical imaging. In *ICDM*, pages 887–892, Dec 2014.

[18] A. Lancichinetti and S. Fortunato. Consensus clustering in complex networks. *Scientific reports*, 2, 2012.

[19] M. Meila and J. Shi. A random walks view of spectral segmentation. In *AISTATS*, 2001.

[20] D. Niu, J. G. Dy, , and M. I. Jordan. Iterative discovery of multiple alternativeclustering views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1340–1353, 2014.

[21] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, second edition, 2006.

[22] Z. Qi and I. Davidson. A principled and flexible framework for finding alternative clusterings. In *ACM SIGKDD*, pages 717–726. ACM, 2009.

[23] K. Ramamohanarao, J. Bailey, and H. Fan. Efficient mining of contrast patterns and their applications to classification. In *Intelligent Sensing and Information Processing, 2005. ICISIP 2005. Third International Conference on*, pages 39–47. IEEE, 2005.

[24] V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML workshop on learning with multiple views*, pages 74–79. Citeseer, 2005.

[25] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003.

[26] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[27] Y. Wakabayashi. The complexity of computing medians of relations. *Resenhas do Instituto de Matemática e Estatística da Universidade de São Paulo*, 3(3):323–349, 1998.

[28] X. Wang, B. Qian, and I. Davidson. On constrained spectral clustering and its applications. *Data Mining and Knowledge Discovery*, 28(1):1–30, 2014.

[29] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In *ICDM*, pages 721–724. IEEE, 2002.

[30] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, pages 1601–1608, 2004.

[31] D. Zhou and C. J. Burges. Spectral clustering and transductive learning with multiple views. In *ICML*, pages 1159–1166. ACM, 2007.