Clustering by Shift

Morteza Haghir Chehreghani
NAVER LABS Europe
(formerly known as Xerox Research Centre Europe - XRCE)
morteza.chehreghani@naverlabs.com

***Abstract*—In order to yield a more balanced partitioning, we investigate the use of additive regularizations for the *Min Cut* cost function, instead of normalization. In particular, we study the case where the regularization term is the sum of the squared size of the clusters, which then leads to shifting (adaptively) the pairwise similarities. We study the connection of such a model with *Correlation Clustering* and then propose an efficient *local search* optimization algorithm to solve the new clustering problem. Finally, we demonstrate the superior performance of our method by extensive experiments on different datasets.**

I. Introduction

Given a set of objects, clustering is concerned with grouping them in such a way that objects of the same group are more similar to each other (according to a predefined similarity measure), compared to those in different groups. This task plays a fundamental role in various data analytics applications. Examples are image segmentation (to detect the objects), document clustering (for the purpose of document organization, topic identification or efficient information retrieval), data compression, and analysis of (e.g., transportation and social) networks and graphs. Clustering itself is not a specific method, rather it is a general machine learning task to be addressed. The task can be solved via several algorithms that differ significantly in the way they define the notion of clusters and the way they extract them. The concept of clustering is originated from anthropology and then was used in psychology [1], [28], in particular for trait theory classification in personality psychology [5].

A wide range of clustering methods introduce a cost function whose minimization yields a clustering solution. *K-means* is a common cost function which is defined by the within-cluster sum of squared distances from the means [20]. The data can be represented by a graph, whose nodes represent the objects and the edge weights are the pairwise similarities between the objects. Then, a wide range of different graph partitioning methods can be applied to produce the clusters. Arguably, the most basic graph-based model is the *Min Cut* (Minimum K-Cut) cost function, in which the goal is to partition the graph into exactly K connected components (clusters) such that the sum of the inter-clusters edge weights is minimal. As we will see, the *Min Cut* cost function often yields separating singleton clusters, in particular when the clusters have diverse densities. To overcome such a problem, several clustering models normalize the *Min Cut* clusters to render more balanced clusters. For example, they propose to normalize the *Min Cut* clusters by the size of the clusters (*Ratio Assoc* [13] and *Ratio Cut* [6]) or the degree of the clusters (*Normalized Cut* [26]). While most of the cost functions assume a nonnegative matrix of pairwise similarities as input, *Correlation Clustering* assumes that the similarities can be negative as well. This model was first introduced on graphs with only $+1$ and -1 edge weights [2], and then it was generalized to graphs with arbitrary positive and negative edge weights [8].

Optimizing such cost functions is often NP-hard [2], [8], [26]. Therefore, the optimal solution should be approximated in some way, e.g., via eigenvector analysis of the respective Laplacian matrix. In this context, *Spectral Clustering* [21], [27] first computes a low-dimensional embedding by the bottom eigenvectors of the Laplacian matrix and then applies K-means to these vectors to yield the final clusters. More recently, instead of embedding the similarities into a K-dimensional space, Power Iteration Clustering (PIC) [17] computes an eigenvalue-weighted combination of all eigenvectors of the normalized similarity matrix via early stopping of the power iteration procedure. P-Spectral Clustering (PSC) [3], [12] develops a non-linear generalization of the Laplacian and then, based on its second eigenvector, iteratively splits the clusters. An alternative graph-based clustering approach has been developed in the context of discrete time dynamical systems and evolutionary game theory which is based on performing replicator dynamics [18], [22], [23]. *Dominant Set Clustering* (DSC) [23] iteratively peels off a cluster by performing a replicator dynamics until its convergence. Then, [18] proposes the two shrink and expansion steps for large datasets with many small and dense clusters. The method in [7] suggests to analyze the trajectory of replicator dynamics in order to accelerate the appearance of clusters. Another method in [4], called *InImDyn*, replaces replicator dynamics by a population dynamics originated from the analogy to infection and immunization processes in a population of players.

In this paper, we investigate adding the regularization terms to the *Min Cut* cost function, in order to avoid creation of singleton sets of clusters. More specifically, we consider the case where the regularization is the sum of the squared size of the clusters, weighted by parameter α. We show that this regularization leads to a simple shift transformation of the input, i.e., subtracting the pairwise similarities by α, which provides a straightforward quadratic model. Such a shift might render some pairwise similarities to be negative, thus, we study the connection to *Correlation Clustering*, another model which works on both positive and negative similarities, and show the

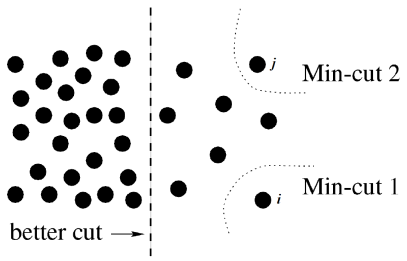

Fig. 1. The *Min Cut* cost function has a bias to split small (singleton) sets of objects. The figure has been adapted from [26].

equivalence of these two models. However, our model provides a principled approach to deduce such negative edge weights (adaptively). Thereafter, we develop an efficient optimization method based on *local search* to solve the new optimization problem. Finally, we perform extensive experiments on several UCI datasets to show the superior performance of our method compared to the alternatives.

II. Shift of Pairwise Similarities for Clustering

The data is given by a set of n objects $\mathbf{O} = \{1, ..., n\}$ and the corresponding matrix of pairwise similarities $\mathbf{X} = \{\mathbf{X}_{ij}\}, \forall i, j \in \mathbf{O}$. Thus, the data can be represented by (an undirected) graph $\mathcal{G}(\mathbf{O}, \mathbf{X})$, where the objects $\mathbf{O}$ constitute the nodes of the graph and $\mathbf{X}_{ij}$ represents the weight of the edge between i and j. Then, the goal is to partition the objects (the graph) into K coherent groups which are distinguishable from each other. The clustering solution is encoded in $\mathbf{c} \in \{1, ..., K\}^n$, i.e., $\mathbf{c}_i$ indicates the cluster label of the i^{th} object. $\mathcal{C}$ denotes the space of all different clustering solutions.

Different graph-based clustering methods often consider the *Min Cut* cost function as a base model which is defined by

$$R^{MC}(c, \mathbf{X}) \quad = \quad \sum_{k=1}^{K} \sum_{\substack{k'=1,\\ k' \neq k}}^{K} \sum_{i \in \mathbf{O}_k} \sum_{j \in \mathbf{O}_{k'}} \mathbf{X}_{ij} \,, \qquad (1)$$

This cost function has a tendency to split small sets of objects, since the cost increases with the number of inter-cluster edge weights, i.e., the edges connecting the different clusters. Figure 1 illustrates such a situation for two clusters [26]. We assume that the edge weights are inversely proportional to the distances between the objects. It is observed that *Min Cut* favors splitting objects i or j, instead of performing a more balanced split. In fact, any cut that splits one of the objects on the right half will yield a smaller cost than the cut that partitions the objects into the left and right halves. This issue is particularly problematic when the intra-cluster edge weights are heterogeneous among different clusters. Thus, several methods propose to normalize the *Min Cut* clusters by a cluster depending factor, e.g., the size of clusters (*Ratio Assoc* [13] and *Ratio Cut* [6]) or the degree of clusters (*Normalized Cut* [26]).

We investigate an alternative approach to yield the occurrence of more balanced clusters. Instead of normalizing (dividing) the *Min Cut* cost function by a cluster-dependent function, we propose to add such a regularization to the original cost function, i.e.,

$$R^{new}(\mathbf{c}, \mathbf{X}, \alpha) \quad = \quad R^{MC}(\mathbf{c}, \mathbf{X}) + \alpha \, . \, r(\mathbf{c}, \mathbf{X}), \qquad (2)$$

where $r(\mathbf{c}, \mathbf{X})$ indicates the regularization. Note that this formulation involves the two free choices α and $r(\mathbf{c}, \mathbf{X})$, thereby, it yields a richer family of alternative models. We particularly focus on the case where $r(\mathbf{c}, \mathbf{X})$ is the sum of the squared size of the clusters[1], i.e.,

$$R^{new}(\mathbf{c}, \mathbf{X}) \quad = \quad R^{MC}(\mathbf{c}, \mathbf{X}) + \alpha \sum_{k=1}^{K} |\mathbf{O}_k|^2 \, . \qquad (3)$$

Thereby, i) if $\alpha < 0$, then the term $\alpha \sum_{k=1}^{K} |\mathbf{O}_k|^2$ is minimal when only the singleton clusters (objects) are separated. Thus, this choice does not help to avoid occurrence of singleton clusters, rather, it facilitates. ii) If $\alpha > 0$, then $\alpha \sum_{k=1}^{K} |\mathbf{O}_k|^2$ is minimal for balanced clusters, i.e., when $|\mathbf{O}_k| = n/K \;, \forall k \in \{1, ..., K\}$. This leads to equalize the size of clusters.

The cost function in Eq. 3 can be further written as

$$\begin{aligned}
R^{new}(\mathbf{c}, \mathbf{X}, \alpha) &= R^{MC}(\mathbf{c}, \mathbf{X}) + \alpha \sum_{k=1}^{K} |\mathbf{O}_k|^2 \\
&= \sum_{k=1}^{K} \sum_{k' \neq k}^{K} \sum_{i \in \mathbf{O}_k} \sum_{j \in \mathbf{O}_{k'}} \mathbf{X}_{ij} + \sum_{k=1}^{K} \sum_{i,j \in \mathbf{O}_k} \mathbf{X}_{ij} \\
&\quad - \sum_{k=1}^{K} \sum_{i,j \in \mathbf{O}_k} \mathbf{X}_{ij} + \sum_{k=1}^{K} \sum_{i,j \in \mathbf{O}_k} \alpha \\
&= \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{i \in \mathbf{O}_k} \sum_{j \in \mathbf{O}_{k'}} \mathbf{X}_{ij} - \sum_{k=1}^{K} \sum_{i,j \in \mathbf{O}_k} (\mathbf{X}_{ij} - \alpha) \\
&= \underbrace{\sum_{i,j \in \mathbf{O}} \mathbf{X}_{ij}}_{constant} - \sum_{k=1}^{K} \sum_{i,j \in \mathbf{O}_k} (\mathbf{X}_{ij} - \alpha) \\
&\equiv - \sum_{k=1}^{K} \sum_{i,j \in \mathbf{O}_k} (\mathbf{X}_{ij} - \alpha) \, . \qquad (4)
\end{aligned}$$

Therefore, we obtain

$$\boxed{R^{SMC}(\mathbf{c}, \mathbf{X}, \alpha) = - \textstyle\sum_{k=1}^{K} \sum_{i,j \in \mathbf{O}_k} (\mathbf{X}_{ij} - \alpha).}$$

Thus, we introduce a shifted variant of *Min Cut* cost function (called *Shifted Min Cut*), wherein all pairwise similarities are subtracted by a positive parameter α, such that some of the pairwise similarities might be negative.

This formulation provides a rich family of alternative clustering models where different regularizations are induced by different values of α. However, choosing a very large α can lead to equalizing the size of the clusters that are inherently very unbalanced in size. For example, consider the dataset

[1]The *Min Cut* cost function is quadratic with respect to the number of edges, therefore, to be consistent, we use the squared form of the cluster cardinalities.

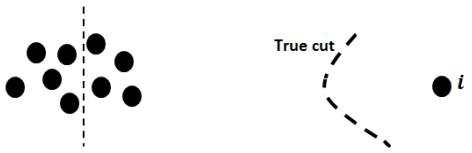

Fig. 2. The impact of the shift parameter α on the results of the *Shifted Min Cut* cost function. A very large α might yield splitting large clusters, instead of separating true small clusters.

shown in Figure 2. We assume that the edge weights are inversely proportional to the pairwise distances. Then, we subtract all pairwise similarities by a very large number. Therefore, the pairwise similarities become very large but negative numbers which renders the *Shifted Min Cut* model to produce equal-size clusters, even though a correct cut should separate only the object i from the rest. Thus, in practice one needs to examine different values of α, and choose the one that yields the best results, or is preferred by the user. However, this procedure might be computationally expensive, and, moreover, the user might not be able to validate the correct solution among many different alternatives, due to lack of enough prior knowledge, supervision or side information. For this reason, we propose a particular shift of pairwise similarities which takes the connectivity of the objects into account and does not require fixing any free parameter.

Adaptive shift of pairwise similarities: Different pairwise similarities might need different shifts, depending on the type and the density of the clusters that the respective nodes belong to. A reasonable approach is to shift the pairwise similarity $\mathbf{X}_{ij}$ between i and j adaptively with respect to the similarities between i and all the other objects and as well as the similarities between j and the other objects. For this purpose, we shift $\mathbf{X}_{ij}$ such that the sum of the pairwise similarities between i and all the other objects becomes zero, and the same holds for j too. Therefore, the new shifted similarity $\mathbf{S}_{ij}$ is obtained by

$$\mathbf{S}_{ij} = \mathbf{X}_{ij} - \frac{1}{n}\sum_{p=1}^{n}\mathbf{X}_{ip} - \frac{1}{n}\sum_{p=1}^{n}\mathbf{X}_{pj} + \frac{1}{n^2}\sum_{p=1}^{n}\sum_{q=1}^{n}\mathbf{X}_{pq}\,. \quad (5)$$

It is easy to check that $\mathbf{S}$ is symmetric, provided that $\mathbf{X}$ is symmetric. It can be shown that sum of the rows and the columns of $\mathbf{S}$ are equal to zero. For example, for a fixed row i we have

$$\begin{aligned}\sum\nolimits_{j=1}^{n}\mathbf{S}_{ij} &= \sum_{j=1}^{n}\mathbf{X}_{ij} - \frac{1}{n}\sum_{j=1}^{n}\sum_{p=1}^{n}\mathbf{X}_{ip} \\ &\quad - \frac{1}{n}\sum_{j=1}^{n}\sum_{p=1}^{n}\mathbf{X}_{pj} + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{p=1}^{n}\sum_{q=1}^{n}\mathbf{X}_{pq} \\ &= \sum_{j=1}^{n}\mathbf{X}_{ij} - \frac{n}{n}\sum_{p=1}^{n}\mathbf{X}_{ip} - \frac{1}{n}\sum_{j=1}^{n}\sum_{p=1}^{n}\mathbf{X}_{pj} + \frac{n}{n^2}\sum_{p=1}^{n}\sum_{q=1}^{n}\mathbf{X}_{pq} \\ &= 0 + 0\,. \end{aligned} \quad (6)$$

The adaptive shift in Eq. 5 can be written in matrix form as

$$\mathbf{S} = \mathbf{T}\mathbf{X}\mathbf{T}\,, \quad (7)$$

where the $n \times n$ matrix $\mathbf{T}$ is defined by $\mathbf{T} = \mathbf{I}_n - \frac{1}{n}\mathbf{U}$. All elements of the $n \times n$ matrix $\mathbf{U}$ are 1.

According to Eq. 4, the new cost function is written by

$$\begin{aligned} R^{SMC}(\mathbf{c},\mathbf{S}) &= -\sum_{k=1}^{K}\sum_{i,j\in\mathbf{O}_k}\mathbf{S}_{ij} & (8)\\ &\equiv \sum_{k=1}^{K}\sum_{k'\neq k}^{K}\sum_{i\in\mathbf{O}_k}\sum_{j\in\mathbf{O}_{k'}}\mathbf{S}_{ij}\,. & (9)\end{aligned}$$

III. Relation to Correlation Clustering

Correlation Clustering partitions a graph with positive and negative edge weights. The cost function sums the disagreements, i.e., the sum of negative intra-cluster edge weights plus the sum of positive inter-cluster edge weights. For a fixed K, the *Correlation Clustering* cost function can be written as

$$\begin{aligned} R^{CC}(c,\mathbf{X}) &= \frac{1}{2}\sum_{k=1}^{K}\sum_{i,j\in\mathbf{O}_k}(|\mathbf{X}_{ij}| - \mathbf{X}_{ij}) \\ &\quad + \frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{k'=1,\\k'\neq k}}^{K}\sum_{i\in\mathbf{O}_k}\sum_{j\in\mathbf{O}_{k'}}(|\mathbf{X}_{ij}| + \mathbf{X}_{ij}). \end{aligned} \quad (10)$$

The first term (called a) sums the intra-cluster negative edge weights, whereas the second term (called b) sums the inter-cluster positive edge weights. We separately expand each term.

$$\begin{aligned} a =& \frac{1}{2}\sum_{k=1}^{K}\sum_{i,j\in\mathbf{O}_k}|\mathbf{X}_{ij}| - \frac{1}{2}\sum_{k=1}^{K}\sum_{i,j\in\mathbf{O}_k}\mathbf{X}_{ij} \\ =& \frac{1}{2}\sum_{k=1}^{K}\sum_{i,j\in\mathbf{O}_k}|\mathbf{X}_{ij}| - \underbrace{\frac{1}{2}\sum_{k=1}^{K}\sum_{k'=1}^{K}\sum_{i\in\mathbf{O}_k}\sum_{j\in\mathbf{O}_{k'}}\mathbf{X}_{ij}}_{constant} \\ &+ \frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{k'=1,\\k'\neq k}}^{K}\sum_{i\in\mathbf{O}_k}\sum_{j\in\mathbf{O}_{k'}}\mathbf{X}_{ij}\,. \end{aligned} \quad (11)$$

Similarly, we expand term b as

$$b = \frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{k'=1,\\k'\neq k}}^{K}\sum_{i\in\mathbf{O}_k}\sum_{j\in\mathbf{O}_{k'}}|\mathbf{X}_{ij}| + \frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{k'=1,\\k'\neq k}}^{K}\sum_{i\in\mathbf{O}_k}\sum_{j\in\mathbf{O}_{k'}}\mathbf{X}_{ij}\,. \quad (12)$$

Then, by summing a and b we obtain

$$\begin{aligned} R^{CC}(c,\mathbf{X}) &= constant + \sum_{k=1}^{K}\sum_{\substack{k'=1,\\k'\neq k}}^{K}\sum_{i\in\mathbf{O}_k}\sum_{j\in\mathbf{O}_{k'}}\mathbf{X}_{ij} \\ &\quad + \underbrace{\frac{1}{2}\sum_{k=1}^{K}\sum_{i,j\in\mathbf{O}_k}|\mathbf{X}_{ij}| + \frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{k'=1,\\k'\neq k}}^{K}\sum_{i\in\mathbf{O}_k}\sum_{j\in\mathbf{O}_{k'}}|\mathbf{X}_{ij}|}_{constant} \\ &\equiv R^{MC}(\mathbf{c},\mathbf{X}). \end{aligned} \quad (13)$$

Thus, *Correlation Clustering* and *Min Cut* are equivalent cost functions, i.e., 1. The cost functions share the same optimal solution, i.e., $\arg\min_{\mathbf{c}} R^{MC}(\mathbf{c},\mathbf{X}) = \arg\min_{\mathbf{c}} R^{CC}(\mathbf{c},\mathbf{X})$. 2. The costs differences are the same, i.e., $\forall \mathbf{c} \in \mathcal{C} : R^{MC}(\mathbf{c},\mathbf{X}) - \min_{\mathbf{c}} R^{MC}(\mathbf{c},\mathbf{X}) = R^{CC}(\mathbf{c},\mathbf{X}) - \min_{\mathbf{c}} R^{CC}(\mathbf{c},\mathbf{X})$.

Thus, *Correlation Clustering*, similar to *Shifted Min Cut*, is a variant of the *Min Cut* model which deals with both negative and positive edge weights. However, there are fundamental differences between these two models: i) *Correlation Clustering* assumes that the matrix of pairwise positive and negative similarities is given (which might be nontrivial), whereas *Shifted Min Cut* proposes a principled way to yield clustering of positive and negative similarities via regularizing the base *Min Cut* model. Thus, *Shifted Min Cut* provides an explicit and straightforward interpretation of the clustering problem. ii) The form of the *Shifted Min Cut* cost function expressed in Eq. 4 (or Eq. 8) provides an efficient function evaluation (e.g., for optimization) compared to the *Correlation Clustering* cost function in Eq. 10 or the base *Min Cut* cost function in Eq. 1. The models in Eqs 1 and 10 are quadratic with respect to K, the number of clusters, whereas the model in Eq. 4 is linear.

IV. Optimization of *Shifted Min Cut*

Finding the optimal solution of the standard *Min Cut* model with non-negative edge weights, i.e., when $\mathbf{X}_{ij} \geq 0, \forall i,j$, is well-studied, for which there exist several polynomial time algorithms, e.g., $\mathcal{O}(n^4)$ [11] and $\mathcal{O}(n^2 \log^3 n)$ [15]. However, finding the optimal solution of the *Shifted Min Cut* cost function, wherein some edge weights are negative, is NP-hard [2], [8] and even is APX-hard [8]. Therefore, we develop a *local search* method which computes a local minimum of the cost function in Eq. 8. The effectiveness of such a greedy strategy is well studied for different clustering cost functions, e.g., K-means [20], kernel K-means [25] and in particular several graph partitioning methods [9], [10].[2] In this approach, we start with a random clustering solution and then we iteratively assign each object to the cluster that yields a maximal reduction in the cost function. We repeat this procedure until no further change of assignments is achieved during a complete round of investigation of the objects, i.e., then a local optimal solution is attained. At each iteration of the aforementioned procedure, one needs to evaluate the cost of assigning every object to each of the clusters. The cost function is quadratic, thus a single evaluation might take $\mathcal{O}(Kn^2)$ runtime. Thereby, if the local search converges after t iterations, then, the total runtime will be $\mathcal{O}(tKn^3)$ for n objects, which might be computationally expensive.

However, we do not need to recalculate the cost function for every individual evaluation. Let $R^{SMC}(\mathbf{c}_{o\to l},\mathbf{S})$ denote the cost of the clustering solution $\mathbf{c}$ wherein object o is assigned to cluster l. At each step of the local search algorithm, we need to evaluate the cost $R^{SMC}(\mathbf{c}_{o\to l'},\mathbf{S}), l' \neq l$ given $R^{SMC}(\mathbf{c}_{o\to l},\mathbf{S})$.

The cost function $R^{SMC}(\mathbf{c}_{o\to l},\mathbf{S})$ is written by

$$R^{SMC}(\mathbf{c}_{o\to l},\mathbf{S}) = -\sum_{k=1}^{K} \sum_{\substack{i,j\in\mathbf{O}_k \\ i,j\neq o}} \mathbf{S}_{ij} - \sum_{\substack{i\in\mathbf{O}_l \\ i\neq o}} (\mathbf{S}_{io}+\mathbf{S}_{oi}) - \mathbf{S}_{oo}\,. \tag{14}$$

Similarly, the cost $R^{SMC}(\mathbf{c}_{o\to l'},\mathbf{S}), l' \neq l$ is obtained by

$$\begin{aligned} R^{SMC}(\mathbf{c}_{o\to l'},\mathbf{S}) &= -\sum_{k=1}^{K} \sum_{\substack{i,j\in\mathbf{O}_k \\ i,j\neq o}} \mathbf{S}_{ij} - \sum_{\substack{i\in\mathbf{O}_{l'} \\ i\neq o}} (\mathbf{S}_{io}+\mathbf{S}_{oi}) - \mathbf{S}_{oo} \\ &= R^{SMC}(\mathbf{c}_{o\to l},\mathbf{S}) + \sum_{\substack{i\in\mathbf{O}_l \\ i\neq o}} (\mathbf{S}_{io}+\mathbf{S}_{oi}) - \sum_{\substack{i\in\mathbf{O}_{l'} \\ i\neq o}} (\mathbf{S}_{io}+\mathbf{S}_{oi}). \end{aligned} \tag{15}$$

Thus, given $R^{SMC}(\mathbf{c}_{o\to l},\mathbf{S})$ the runtime of a new evaluation of the cost function $R^{SMC}(\mathbf{c}_{o\to l'},\mathbf{S})$ is $\mathcal{O}(n)$. Hence, the total runtime of the local search method will be $\mathcal{O}(tn^2)$. Therefore, at the beginning, we compute a random initial solution, wherein each object is assigned randomly to one of the K clusters, and compute the respective cost. At each iteration, we use Eq. 15 to investigate the cost of assigning an object to the other clusters than the current one. Then, we assign the object to the cluster that yields a maximal reduction in the cost. We might repeat the local search algorithm with several random initializations and at end, choose a solution with a minimal cost. Note that even the efficient evaluation and optimization of the model variants in Eq. 1 and Eq. 10 would yield $\mathcal{O}(tKn^2)$ total runtime, i.e., K times slower than the variant derived in Eq. 4 (and expressed in Eq. 8). Notice this technique can be employed with other optimization or inference methods as well, such as MCMC methods and simulated annealing.

V. Experiments

We empirically investigate the performance of our clustering method and compare the results against several alternatives. We perform the experiments under identical computational settings on a core i7-4600U Intel machine with 2.7 GHz CPU and 8.00 GB internal memory.

Data: We perform our experiments on several UCI datasets [16], chosen from different domains and contexts. 1. *Breast Tissue*: contains 106 electrical impedance measurements of the breast tissue samples in 6 types each with 10 features. 2. *Cloud*: consists of 1024 vectors, where each vector includes 10 parameters. 3. *Ecoli*: a biological data on the cellular localization sites of 7 types of proteins which includes 336 samples. 4. *Forest Type Mapping*: a remote sensing dataset of 326 samples with 27 attributes collected from forests in Japan grouped in 7 different categories. 5. *Heart*: dataset of heart disease (absence or presence) that involves 270 instances and 13 attributes. 6. *Lung Cancer*: high-dimensional lung cancer data with 32 instances and 56 features. 7. *Parkinsons*: contains 197 biomedical voice measurements 23 attributes.

[2]Consistently, with *Correlation Clustering* we observe a significantly better performance of the local search algorithm compared to approximation schemes such as those proposed in [2], [8].

8. *Pima Indians Diabetes*: the data of 768 female patients from Pima Indian heritage with 8 attributes. 9. *SPECTF*: describes SPECTF images with 44 attributes about the heart of 267 patients. 10. *Statlog ACA (Australian Credit Approval)*: contains information of 690 credit card applications each described with 14 features. 11. *Teaching Assistant*: consist of evaluations of teaching performance over 5 semesters of 151 teaching assistant assignments. 12 *User Knowledge Modeling*: contains the 403 students' knowledge status on Electrical DC Machines with 5 attributes and grouped in 4 categories.

In these datasets, the objects are represented by vectors. Thus, to obtain the pairwise similarity matrix $\mathbf{X}$, we first compute the pairwise squared Euclidean distances between the vectors and obtain matrix $\mathbf{D}$. Then, as proposed in [7], we convert the pairwise distances $\mathbf{D}$ to the similarity matrix $\mathbf{X}$ via $\mathbf{X}_{ij} = \max(\mathbf{D}) - \mathbf{D}_{ij} + \min(\mathbf{D})$, where the $\max(.)$ and $\min(.)$ operations respectively give the maximum and the minimum of the elements of $\mathbf{D}$. An alternative transformation is an exponential function in the form of $\mathbf{S}_{ij} = \exp(-\frac{\mathbf{X}_{ij}}{\sigma^2})$, which requires fixing the free parameter σ in advance. However, this task is nontrivial in unsupervised learning and the appropriate values of σ fall in a very narrow range [19].

Methods: We compare our method, called *Shifted Min Cut* against several competitive methods in the literature. We consider the following methods: i) Dominant Set Clustering (DSC), ii) InImDyn, iii) P-Spectral Clustering (PSC), iv) Gaussian Mixture Models (GMM), v) K-means, vi) Power Iteration Clustering (PIC), vii) and Spectral Clustering (SC). We run each method 100 times with different random initializations and choose the best solution in terms of the cost or likelihood. With the GMM method, we obtain the probabilistic assignment of the objects to the clusters. Then, we assign each object to the most probable cluster.

Evaluation criteria: We have access to the ground truth solutions of the datasets. Thus, we compare the true labels and the estimated solutions to investigate quantitatively the performance of each method. For this purpose, we measure three evaluation criteria: i) adjusted Mutual Information [29]: the mutual information between the two estimated and true clustering solutions, ii) adjusted Rand score [14]: the similarity between the solutions, and iii) V-measure [24]: the harmonic mean of homogeneity and completeness. We compute the adjusted variant of these criteria, i.e., they yield zero for random solutions.

Results: Tables I, II and III show the results of different clustering methods on the UCI datasets respectively with respect to the Mutual Information criterion, the Rand score and the V-measure. We observe that for most of the datasets, *Shifted Min Cut* yields the best scores. In the cases that the method is not the best, it is usually among top choices. DSC and InImDyn perform very similarly, consistent to the results in [4]. PIC works well in the cases that there are few clusters in the dataset. The reason is that it computes an one-dimensional embedding of the data and then performs K-means. However, such an embedding might confuse some clusters when there exist many of them in the dataset. PSC is significantly slower than the other methods and also yields suboptimal results, as reported by several previous studies as well.

VI. Conclusion

This paper advocates an alternative approach for regularizing the *Min Cut* cost function in order to avoid the appearance of singleton clusters, where the regularization term is added to the cost function, instead of dividing the *Min Cut* clusters by a cluster dependent factor. We, in particular, studied the case where the regularization term leads to subtracting the pairwise similarities by the regularization factor. Then, we only need to apply the base *Min Cut* model, but on the (adaptively) shifted similarities instead of the original data. In the following, we developed an efficient *local search* algorithm to optimize (locally) the *Shifted Min Cut* cost function. At the end, we performed extensive experiments on several UCI and real-world datasets to demonstrate the better performance of *Shifted Min Cut* according to different evaluation criteria.

References

[1] K. D. Bailey. *Numerical Taxonomy and Cluster Analysis*. SAGE Publications, 1994.

[2] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56(1-3):89–113, 2004.

[3] T. Bühler and M. Hein. Spectral clustering based on the graph p-laplacian. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 81–88. ACM, 2009.

[4] S. R. Bulò, M. Pelillo, and I. M. Bomze. Graph-based quadratic optimization: A fast evolutionary approach. *Computer Vision and Image Understanding*, 115(7):984–995, 2011.

[5] R. B. Cattell. The description of personality: basic traits resolved into clusters. *The Journal of Abnormal and Social Psychology*, 38(4):476–506, 1943.

[6] P. K. Chan, M. D. F. Schlag, and J. Y. Zien. Spectral k-way ratio-cut partitioning and clustering. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 13(9):1088–1096, 1994.

[7] M. H. Chehreghani. Adaptive trajectory analysis of replicator dynamics for data clustering. *Machine Learning*, 104(2-3):271–289, 2016.

[8] E. D. Demaine, D. Emanuel, A. Fiat, and N. Immorlica. Correlation clustering in general weighted graphs. *Theor. Comput. Sci.*, 361(2-3):172–187, 2006.

[9] I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 551–556. ACM, 2004.

[10] I. S. Dhillon, Y. Guan, and B. Kulis. A unified view of kernel k-means, spectral clustering and graph cuts. Technical Report TR-04-25, 2005.

[11] O. Goldschmidt and D. S. Hochbaum. A polynomial algorithm for the k-cut problem for fixed k. *Mathematics of Operations Research*, 19(1):24–37, 1994.

[12] M. Hein and T. Bühler. An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA. In *Advances in Neural Information Processing Systems 23*, pages 847–855, 2010.

[13] T. Hofmann and J. M. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(1):1–14, 1997.

[14] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[15] D. R. Karger and C. Stein. A new approach to the minimum cut problem. *J. ACM*, 43(4):601–640, 1996.

[16] M. Lichman. UCI machine learning repository, 2013.

[17] F. Lin and W. W. Cohen. Power iteration clustering. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 655–662, 2010.

[18] H. Liu, L. J. Latecki, and S. Yan. Fast detection of dense subgraphs with iterative shrinking and expansion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(9):2131–2142, 2013.

TABLE I

PERFORMANCE OF DIFFERENT METHODS WITH RESPECT TO THE ADJUSTED MUTUAL INFORMATION CRITERION. *Shifted Min Cut* YIELDS THE BEST RESULTS IN MOST OF THE CASES.

dataset	ShiftedMinCut	DSC	InImDyn	PSC	GMM	K-means	PIC	SP
Breast Tissue	0.4196	0.4305	0.4196	0.3606	0.3276	0.1809	0.4123	**0.4507**
Cloud	**1.0000**	0.3812	0.3543	0.3098	0.8511	0.3056	0.8597	0.8406
Ecoli	0.5414	0.4731	0.4368	0.5074	**0.5800**	0.5685	0.0542	0.4743
Forest Type	0.4352	0.2960	0.3109	0.2704	0.3877	**0.5197**	0.3875	0.3163
Heart	**0.1570**	0.0698	0.0602	0.0594	0.0813	0.0813	0.0509	0.1078
Lung Cancer	0.2362	0.0850	0.0859	0.0713	0.1684	0.1997	0.0380	**0.2473**
Parkinsons	**0.1957**	0.0738	0.0761	0.0511	0.0484	0.0136	0.0153	0.1631
Pima Indians Diabetes	0.1178	0.0561	0.0533	0.0368	0.0003	0.0257	0.1226	0.1200
SPECTF	**0.1570**	0.0698	0.0602	0.0419	0.0813	0.0813	0.0509	0.1078
Statlog ACA	**0.3907**	0.1607	0.1498	0.1392	0.0038	0.0038	0.3570	0.3683
Teaching Assistant	**0.1041**	0.0268	0.0357	0.0123	0.0353	0.0130	0.0470	0.0123
User Knowledge Modeling	0.2926	0.1107	0.1198	0.0441	**0.6100**	0.2139	0.2454	0.1169

TABLE II

PERFORMANCE OF DIFFERENT METHODS WITH RESPECT TO THE ADJUSTED RAND SCORE. *Shifted Min Cut* LEADS TO BETTER CLUSTERING SOLUTIONS ON MOST OF THE DATASETS.

dataset	ShiftedMinCut	DSC	InImDyn	PSC	GMM	K-means	PIC	SP
Breast Tissue	**0.3546**	0.2929	0.2907	0.3100	0.2085	0.0943	0.3125	0.3037
Cloud	**1.0000**	0.3573	0.3117	0.2926	0.8991	0.2429	0.9065	0.8899
Ecoli	**0.6801**	0.4068	0.3299	0.5145	0.5574	0.4944	0.0378	0.4132
Forest Type	0.3983	0.2225	0.2426	0.2027	0.3285	**0.4987**	0.3560	0.2454
Heart	**0.0917**	0.0608	0.0449	0.0578	0.0467	0.0337	0.0721	0.0588
Lung Cancer	0.2962	0.0689	0.0664	0.0215	0.1698	0.2294	0.0949	**0.4201**
Parkinsons	**0.1275**	0.0129	0.0360	0.0234	0.0123	0.0162	0.0178	0.0419
Pima Indians Diabetes	0.1535	0.0454	0.0460	0.0381	0.0010	0.0744	**0.1617**	0.1200
SPECTF	**0.0917**	0.0608	0.0429	0.0570	0.0617	0.0617	0.0434	0.0898
Statlog ACA	**0.4913**	0.1485	0.1307	0.1332	0.0022	0.0022	0.4550	0.4710
Teaching Assistant	**0.1170**	0.0112	0.0127	0.0129	0.0220	0.0089	0.0322	0.0129
User Knowledge Modeling	0.2912	0.0713	0.0778	0.0503	**0.5680**	0.1675	0.1449	0.1053

TABLE III

PERFORMANCE OF DIFFERENT ALGORITHMS WITH RESPECT TO THE ADJUSTED V-MEASURE. IN A CONSISTENT WAY TO THE TWO PREVIOUS EVALUATION CRITERIA, THE *Shifted Min Cut* METHOD PROVIDES THE BEST CLUSTERING RESULTS ON MOST OF THE DATASETS.

dataset	ShiftedMinCut	DSC	InImDyn	PSC	GMM	K-means	PIC	SP
Breast Tissue	**0.5563**	0.4914	0.4999	0.4756	0.4389	0.2895	0.5230	0.5142
Cloud	**1.0000**	0.5525	0.5239	0.5116	0.8520	0.3388	0.8605	0.8417
Ecoli	**0.6396**	0.5339	0.5169	0.5203	0.6192	0.6321	0.1139	0.5364
Forest Type	0.4835	0.3591	0.3866	0.2729	0.3937	**0.5279**	0.3982	0.3244
Heart	**0.1788**	0.1186	0.1061	0.0689	0.0896	0.0896	0.0644	0.1131
Lung Cancer	0.2730	0.2175	0.2310	0.1038	0.2030	0.2356	0.1117	**0.2987**
Parkinsons	**0.2196**	0.1255	0.1327	0.0814	0.0130	0.0105	0.0291	0.1798
Pima Indians Diabetes	0.1227	0.0867	0.0838	0.0693	0.0013	0.0295	**0.1276**	0.1200
SPECTF	**0.1788**	0.1186	0.1061	0.0819	0.0896	0.0896	0.0644	0.1131
Statlog ACA	**0.3927**	0.2396	0.2267	0.2068	0.0099	0.0099	0.3632	0.3720
Teaching Assistant	**0.1156**	0.0615	0.0770	0.0253	0.0583	0.0263	0.0606	0.0253
User Knowledge Modeling	0.3384	0.1527	0.1697	0.0573	**0.6191**	0.2278	0.2603	0.1325

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

[20] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[21] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2001.

[22] B. Ng, M. J. McKeown, and R. Abugharbieh. Group replicator dynamics: A novel group-wise evolutionary approach for sparse brain network detection. *IEEE Trans. Med. Imaging*, 31(3):576–585, 2012.

[23] M. Pavan and M. Pelillo. Dominant sets and pairwise clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):167–172, 2007.

[24] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, pages 410–420. ACL, 2007.

[25] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299–1319, 1998.

[26] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[27] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[28] R. Tryon. *Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality*. Edwards brother, Incorporated, lithoprinters and publishers, 1939.

[29] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, 11:2837–2854, 2010.